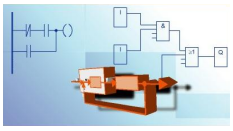
Universiteit Gent
Faculteit Ingenieurswetenschappen
Vakgroep Electrische Energie,
Systemen en Automatisatisering

Berekenbaarheid in statistische hypothesetoetsen en karakterisering van onafhankelijkheid en gerichte beïnvloeding in tijdsreeksen, gebruik makend van Kolmogorov-complexiteit

Computability in statistical hypotheses testing, and characterizations of independence and directed influences in time series using Kolmogorov complexity

Bruno Bauwens



Proefschrift tot het bekomen van de graad
Doctor in de Ingenieurswetenschappen:
Wiskundige Ingenieurstechnieken
Academiejaar 2009-2010

Universiteit Gent
Faculteit Ingenieurswetenschappen
Vakgroep Electrische Energie,
Systemen en Automatisatisering

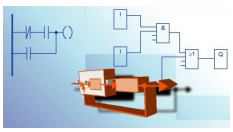
Promotoren: Prof. Dr. Ir. Luc Boullart
Prof. Dr. Patrick Santens

Universiteit Gent
Faculteit Ingenieurswetenschappen

Vakgroep Elektrische Energie, Systemen and Automatisatisering
Technologiepark 913, B-9052 Zwijnaarde, België

Tel.: +32-9-264.55.76
Fax.: +32-9-264.35.82

Dit werk kwam tot stand in het kader van een specialisatiebeurs van het IWT-Vlaanderen (Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen).



Proefschrift tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen:
Wiskundige ingenieurstechnieken
Academiejaar 2009-2010

Dankwoord

dankwoord...

Gent, februari 2010
Bruno Bauwens

Table of Contents

Samenvatting

Een enkelvoudige statistische hypothese is een statistische hypothese waarmee men een unieke probabiliteitsdistributie kan vormen voor alle mogelijke te verwachten data voor een experiment. Significanties en powers zulke statistische testen kunnen nu frequentistisch en objectief geïnterpreteerd worden. Vervolgens geven we een inleiding tot berekenbaarheidstheorie en Kolmogorov complexiteit, enerzijds een tool om efficient combinatorische problemen op te lossen, anderzijds een grootheid die gebruikt wordt om ideale statistische testen te bepalen, en het gebruik van data compressie heuristieken in algoritmes rechtvaardigd.

Twee vragen worden traditioneel beantwoord gebruikmakend van statistische hypothese toetsen:

- Is de data typisch voor een model ?
- Welke van twee modellen verklaren de data het best ?

De eerste vraag wordt beantwoord gebruikmakend van significantietesten. Hierbij wordt een gegeven subset van de mogelijke data gedefinieerd waarvoor volgens de hypothese er slechts een kleine kans bestaat dat één van die data geobserveerd wordt. Wanneer dit gebeurt, wordt besloten dat de data niet typisch is voor het model. Somtesten voor een berekenbare probabiliteitsdistributie vormen een abstract model voor significanties van statistische testen voor een simpele hypothese. De vraag stelt zich voor welke berekenbaarheidsklassen optimale somtesten bestaan, en voor welke berekenbaarheidsklassen de grootste testen bestaan. Het is bekend dat voor beneden-benaderbare sumtesten er zo'n optimaal element bestaat. Het is een open probleem of dat er boven-berekenbare testen bestaan die onbeperkt groter kunnen zijn dan een beneden-berekenbare test.

Veel hypothesen, zoals bijvoorbeeld de algemene hypothese van onafhankelijkheid van variabelen, zijn geen enkelvoudige hypothese, maar samengestelde hypothesen. Semimeasures zijn probabiliteitsdistributies waarvan de som van de probabiliteiten kleiner dan of gelijk aan één kan zijn. Voor een grote groep van samengestelde hypothesen heeft de corresponderende verzameling van beneden-berekenbare semimeasures een universeel element. We argumenteren dat somtesten voor samengestelde hypothesen op een gelijkaardige manier kunnen geïnterpreteerd als frequentistische of objective significancies. Dezelfde vragen kunnen nu gesteld

worden voor sumtesten voor beneden-berekenbare semimeasures als voor berekenbare probabiliteitsdistributies: is er binnen een gegeven berekenbaarheidsklasse een optimale somtest, en welke berekenbaarheidsklasse bevat de grootste somtesten. Er wordt aangetoond dat er beneden berekenbare semimeasures zijn die geen universele beneden-berekenbare somtests hebben, en dat er voor somige semimeasures de boven-berekenbare somtesten de beneden-berekenbare somtesten domineren. Twee specifieke gevallen worden bestudeerd: onafhankelijkheidstesten, en somtesten voor een universele beneden-berekenbare semimeasure.

Onafhankelijkheidstesten worden geïntroduceerd als somtesten $d(x, y)$ voor semimeasures $P(x, y) = Q(x)R(y)$. We tonen aan dat alle beneden-berekenbare somtesten begrensd zijn door een constante, en dat de boven-berekenbare somtesten maximaal kunnen zijn. Vervolgens tonen we aan dat er voor elke boven-benaderbare onafhankelijkheids somtest d er een boven-benaderbare somtest d', en binaire sequenties x, y zijn zodat d geen afhankelijkheden kan vinden tussen x, y, terwijl d' een maximale afhankelijkheid ontdekt. Dit toont aan dat er geen optimale boven-benaderbare somtesten zijn.

Voor boven-berekenbare somtesten voor een universele semimeasure tonen we aan dat ze groter kunnen zijn dan $\log l(x) - O(\log \log l(x))$, maar niet groter dan $\log l(x) + O(\log \log l(x))$. Hieruit zal volgen dat enkele andere bijna-somtesten zoals "coarse sophistication", en de lengte van een minimaal typisch model voor x, niet van bovenuit benaderbaar zijn, zelfs niet wanneer zeer grote fouten toegelaten zijn.

Ook het probleem van modelselectie wordt onderzocht. In veel algoritmen in machinaal leren worden modellen gezocht die naast een bepaald doel te optimaliseren ook een minimale omvang hebben. Het blijkt dat deze modellen robuster zijn. Wanneer het doel is het modelleren van data, wil men daarom enerzijds een model bekomen dat alle regulariteiten van de data bevat, en anderzijds zo klein mogelijk is. De vraag stelt zich of beide criteria in een ideale context terzelfdertijd kunnen voldaan zijn. Wanneer we dit probleem formalizeren aan de hand van algorithmische minimale voldoende statistieken, en algorithmische minimale typische modellen, zullen we aantonen dat dit mogelijk is binnen algorithmische nauwkeurigheid in de beschrijvingslengte van de data. Het is echter niet mogelijk voor heel grote complexe data om een exacte gelijkheid te bekomen. Wanneer we echter zwakke minimale voldoende statistieken introduceren, dan kunnen we aantonen dat ze precies overeenkomen met minimale typische modellen. Bovendien zijn deze steeds equivalent met een initieel segment van de Halting probabiliteit.

Tenslotte bestuderen we de tweede vraag voor statistische hypothese toetsen: welke van de twee hypothesen beschrijft de data het best ? Bij enkelvoudige hypothesen gebeurt dit optimaal door de breuktest. We zullen argumenteren dat breuktesten voor universele semimeasures een alternatief definiëren voor de breuktesten die gebruikt worden om enkelvoudige hypothesen te toetsen.

Om voldoende accuraat verschillende hypothesen te formuleren over onafhankelijkheid en causaliteit, wordt het formalisme van objectieve probabiliteiten verder ontwikkeld gebruikmakend van berekenbaarheidstheorie. Op deze wijze definiëren we verschillende hypothesen voor causale beïnvloeding in tijdsreeksen. Er wordt aangetoond dat alle nodige hypothesen op deze manier gedefiniëerd universele beneden-berekenbare semimeasures definieren. Een tweede manier waarop causale semimeasures kunnen gedefinieerd worden, is door deze op Bayesiaanse manier te associeren vanuit conditionele semimeasures. In de meest algemene vorm krijgen we een sterk verschillende klasse van causale semimeasures. Klasse van causale semimeasures geassocieerd met universele beneden-berekenbare semimeasures is disjunct met de klasse van de universele beneden-berekenbare causale semimeasures. Hoe groot de verschillen tussen semimeasures binnen deze klasse tussen deze klasse en de universele beneden-berekenbare causale semimeasures kunnen is een open vraag. De hypothese testen voor beïnvloeding, gedefiniëerd aan de hand van deze semimeasures, zijn nauw gelinkt met Shannon informatie transfer, en vormen hierom een idealizatie van algoritmen die Shannon informatie transfer schatten. Aan de andere zijde vormen hypothese testen gebaseerd op universele beneden-berekenbare causale semimeasures, een idealizatie voor algoritmen gebaseerd op Granger causaliteit. We definieren het vermoeden dat er een nauw verband is tussen deze twee testen, en dat er dus op theoretisch niveau een nauw verband is tussen algoritmen voor het bepalen van gerichte beïnvloeding in tijdsreeksen aan de hand van Shannon informatie transfer en Granger causaliteit.

Verder zullen we een groot aantal online Kolmogorov complexiteiten definieren die de grootte van deze universele semimeasures benaderen. Hiervoor worden totale online Kolmogorov complexiteiten geïntroduceerd. Deze kunnen substantieel verschillen van de gewone online complexiteiten, voor binare strings die Halting informatie bevatten. Ze stellen ons in staat een decompositie van Kolmogorov complexiteit door te voeren. Verder zijn er nog een aantal varianten van online Kolmogorov complexiteiten waarvan het een open vraag is of die ook decomposities van Kolmogorov complexiteiten definieren. Van gewone online Kolmogorov complexiteiten tonen we aan dat ze niet additief zijn. Dit is een bijzondere eigenschap aangezien die in bepaalde gevallen toelaten om gelijktijdige oorzaken en gevolgen te kunnen onderscheiden, wanneer Halting informatie uitgewisseld wordt.

Summary

In statistics, a simple hypothesis is a hypothesis that implies a probability distribution for all a-priori expected data of an experiment. Hypotheses tests for two such simple hypotheses can now be defined and the corresponding significance and power can be given frequentist and objective interpretations. Subsequently, a basic tutorial on computability and Kolmogorov complexity is given, on one side it is a powerful tool to handle combinatorial problems efficiently in computability theory, on the other side it characterizes many ideal hypotheses tests and justifies the use of data compression heuristics in practical algorithms.

The two statistical hypotheses testing questions are:

- Is the data typical for a given model ?
- Which of two models should be preferred according to the data ?

The first question is typically addressed using significance testing: a set of data is a priori defined, for which there is a low probability that one of its elements will be observed according to the hypotheses. If an element in this set is observed, the hypotheses is rejected. Sumtests relative to computable probability distributions provide an abstract model for the procedure of significance testing. The question rises whether in some computability class there is an optimal sumtest for a computable probability distribution, and which computability class contains the largest elements. It is well known that the class of lower semicomputable sumtests has such an optimal element. The question whether upper semicomputable sumtests can unboundedly exceed lower semicomputable semimeasures is left open.

Many hypotheses, such as the general hypothesis of independence of two observables, are not simple, but composite. A semimeasure is a probability distribution for which the sum of the probabilities may be lower or equal to one. For a large group of composite hypotheses, it is shown that that there exists a universal lower semicomputable semimeasure in the set of corresponding semimeasures. It is argued that sumtests for such a universal semimeasures define significances with a similar frequentist and objective interpretation. The same questions for sumtests relative to computable semimeasures can now be asked for sumtests relative to lower semicomputable semimeasures. Is there within some computability class an optimal sumtest, and for which class is there a largest sumtest ? It is shown

that there exists lower semicomputable semimeasures that have no universal lower semicomputable sumtests, and that there exists lower semicomputable semimeasures for which the upper semicomputable sumtests additively exceed any lower semicomputable sumtests unboundedly. Two specific cases are investigated: independence tests, and sumtests for a universal semimeasure.

Independence tests are introduced as sumtests $d(x, y)$ for semimeasures $P(x, y) = m(x)m(y)$ where m is a universal semimeasure. It is shown that there are no unbounded lower semicomputable independence tests, but that upper semicomputable independence tests can have a maximal value. Furthermore it is shown that for any upper semicomputable test d there are sequences x, y such that d finds almost no dependence between x, y and some upper semicomputable test d' finds an almost maximal dependence.

Sumtests for a universal semimeasures are of very non-trivial nature. It is shown that these can exceed $\log l(x) - O(\log \log l(x))$ but can not be larger than $\log l(x) + O(\log \log l(x))$. This result implies that many interesting functions that are within a logarithmic additive term sumtests for a universal semimeasure, can not be approximated by lower or upper semicomputable functions. This shows that coarse sophistication of x, and the length of a minimal typical model for x can not be approximated by a lower or upper semicomputable function, not even within very large error bounds.

Also, the problem of ideal model selection is investigated from a computability point of view. Many machine learning algorithms select models by that are not only good at solving some predefined task, but are at the same time have a minimal size. It is often observed that such programs are more robust in performance while solving related tasks. When the goal is to extract all regularities out of the data, the question raises whether in theory, both criteria can be satisfied at the same time. This problem can be formalized using algorithmic minimal sufficient statistics, and algorithmic typical models. It is shown that using probabilistic models, these goals can be optimized simultaneously, within logarithmic bounds of the complexity of the data by the same model. For very long and complex data sequences it is shown that an exact equivalence is not possible in general. To solve this issue, weak sufficient statistics are introduced as a variant of sufficient statistics. It is shown that they are equivalent to minimal typical models, and they are both equivalent to an initial segment of the Halting probability.

Finally, the second statistical hypotheses question is addressed: which of two models describes the data best? For simple hypotheses this happens in an optimal accepted way, by ratio tests. It will be argued that ratio tests relative to universal lower semicomputable semimeasures in a hypotheses, can be used to test composite hypotheses.

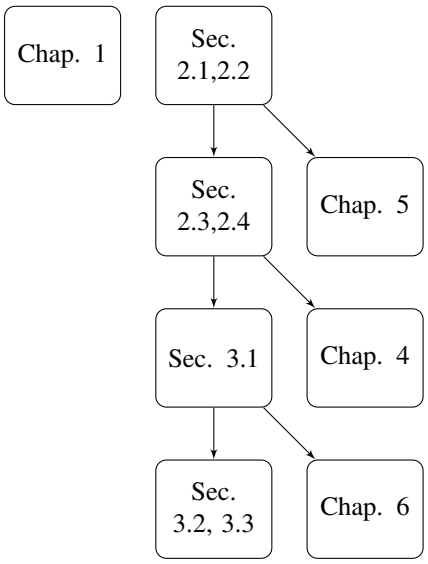
In order to define hypothesis sufficiently rigorous, the objective uncertainty formalism is developed in detail, within the computability framework. Using this

framework several causal and influence-free semimeasures are defined which all have universal elements. With these universal semimeasures several tests for causality and directed influences in time series can be defined. The corresponding sets of semimeasures all have universal elements. Also, another approach of defining causal and influence-free semimeasures is possible by Bayesianally associate a causal semimeasure with a universal semimeasure. In general, these classes of semimeasures differ substantially, however, the the class of causal semimeasures associated with universal lower semicomputable semimeasures will also be considered. This class is conjectured to be disjunct from the class of universal causal lower semicomputable semimeasures. It is an open question how much the semimeasures in this associated class can differ, and how much these semimeasures can differ from the universal lower semicomputable causal semimeasures. The hypotheses that two timeseries are influence-free, based on this class of causal semimeasures, is closely connected to the use of Shannon information transfer, and therefore defines an idealization of algorithms inspired by the concept of Shannon information transfer. On the other side, universal lower semicomputable semimeasures define some idealization of the concept of Granger causality. A tight relation is conjectured between both tests, which also can relate theoretically both approaches.

Furthermore, a large number of online complexities will be defined. A coding theorem is proved, which links the corresponding tests to online Kolmogorov complexities. The online Kolmogorov complexities have a remarkable property that they are non additive in a lineair sense. A variant of the online complexities is introduced, which is called total online complexity, and it is shown that this complexity satisfies such an additivity property. The non additivity of online complexities allows to define in some cases simultaneous “cause” and “consequence” in time series when Halting information is transmitted.

Order of reading

The manuscript is self contained, only relying on basic mathematical notions. The optimal order of reading is the order presented here, however, depending on the background and the interests of the reader, other orders of reading can be preferred. People interested in statistical hypothesis testing and causality, can read Chap. 1, Sec. 2.1, Sec. 2.2, and Chap. 5. These chapters and sections do almost not rely on Kolmogorov complexity. People interested in the details of minimal sufficient statistics and ideal model selection can read Chap. 2 and 3. Additional advised orders of reading are summarized in scheme .



1

Statistical hypotheses testing

Science is about agreeing on how to model
in the world of sensory experience,
the words "fact", and "reality"
are used for other purposes.

Abstract. The chapter provides concepts to motivate the mathematical questions defined in this work. First, scientific modeling, and the role of frequentist and objective probabilities is discussed. This leads to the two main statistical hypotheses questions:

- Is the data typical for the model defined by a hypothesis ?
- Which of two hypotheses are favored by the data ?

The objective formalism of probability defines a set of typical models relative to observed data, and is therefore related to the first question.

An alternate approach to answer the first question for simple hypotheses, is given by sumtests. The definition of sumtests defines an abstract model for significance testing. For more general hypotheses, it is argued that a similar approach can be applied by defining sumtests for universal semimeasures.

To answer the second question, in the case of simple hypotheses, several approaches lead to the conclusion that ratio tests are optimal: uniformly most powerful tests, Occam's razor arguments, and the formalism of subjective Bayesian belief factors. Since the latter is especially useful to interpret statistical hypothesis tests as a tool to find agreement, it is explained in more detail. It is argued that ratio tests for two universal semimeasures provide a

theoretical method for statistically favoring one of two non-simple hypotheses.

Remark that future chapters do not rely on definition or concepts discussed here. A mathematical introduction is given in Chapter 2.

1.1 Probabilistic scientific modeling

Scientific modeling is the process of making logic *rules* for symbols that represent *observables* or properties of observables in specific *contexts*. These rules can be iteratively applied, and combined to reproduce past observations or predict, and control future observations in specific contexts.

Observables may vary within a specific context. Such observables are called *variables*. The resulting logic often has "uncontrolled" variables or "unobservable" variables. Let A, B be observable variables, and let a, b be values for A, B. An example of such a logic rule is: "If $A = a$ is observed then $B = b$".

A special category of logical rules, are rules expressing some uncertainty. Suppose for example that if $A = a$ is observed then $B = b$ is "likely". Within scientific modeling there are several formalisms to express uncertainty. A formalism to express uncertainty must provide rules to:

- add new probabilistic logical rules to a model (induction),
- apply probabilistic rules in a model to predict or control future observations (application).

Frequentistic, and objective formalisms are investigated[1]. Also, a subjective belief formalism expressing uncertainty will be discussed in the following section.

1.1.1 Kolmogorov axioms of probability

In [38], Kolmogorov introduces a set of axioms defining the concept of probability. The axioms follow in a very plausible way from either the frequentist, objective, and subjective formalism of uncertainty, and are stated here for later reference. Let E be a set, and let Ω be a set of subsets of E, which are called random events.

I. Ω is a field of sets[2].

[1] Remark that the so called problem of "foundations of probability" or "interpretations of probability" has a very difficult history, which suffered from several ideological, and political interference [73]. The current situation is that most statisticians ignore this issue, and leave the interpretation of probabilities to the students, without further help. Mathematics students are interested in properties of formalisms, and do not see problems here. However, in applied disciplines most students have an unsatisfied feeling by the weak connection with the world of sensory experience.

[2] Currently, a σ-algebra is used here, a field is a collection of sets closed under relative complementation, finite union, and intersection.

II. $E \in \Omega$

III. To each set $A \in \Omega$, a nonnegative real number $P(A)$ is assigned, which is called *probability* of event A. The corresponding function is called *probability distribution*.

IV. $P(E) = 1$

V. If A and B are disjoint, then

$$P(A \cup B) = P(A) + P(B)$$

He also added a sixth axiom, which is redundant for finite Ω. The Bayesian axiom also appears in the frequentist, and objective formalism.

$$P(A|B) = \frac{P(A \cup B}{P(B)}.$$

1.1.2 Frequentistic formalism of uncertainty

In the frequentist formalism, probabilities are related to observation by frequencies of occurrence, both for induction and application of probabilistic rules. This formalism was initiated by Von Mises [72], and further developed by Kolmogorov in [38]. A translation of the section "The relation to the world of experience" in [38], is available in [73], which is cited here.

The theory of probability is applied to the real world of experience as follows:

1. *Suppose we have a certain system of conditions S, capable of unlimited repetition (*context*).*

2. *We study a fixed circle of phenomena that can arise when the conditions S are realized. In general, these phenomena can come out in different ways in different cases where the conditions are realized. Let E be the set of the different possible variants $\xi_1, \xi_2, \ldots$ of the outcomes of the phenomena. Some of these variants might actually not occur. We include in the set E all the variants we regard a priori as possible.*

3. *If the variant that actually appears when conditions S are realized, belongs to a set A that we define in some way, then we say that the event A has taken place. Example. The system of conditions S consists of flipping a coin twice. The circle of phenomena mentioned in point 2 consists of the appearance, on each flip, of head or tails. It follows that there are four possible variants (elementary events), namely headsheads, headstails, tailsheads, tailstails.*

Consider the event A that there is a repetition. This event consists of the first and fourth elementary events. Every event can similarly be regarded as a set of elementary events.

4. *Under certain conditions, that we will not go into further here, we may assume that an event A that does or does not occur under conditions S is assigned a real number $P(A)$ with the following properties: A. One can be practically certain that if the system of conditions S is repeated a large number of times, n, and the event A occurs m times, then the ratio m/n will differ only slightly from $P(A)$. B. If $P(A)$ is very small, then one can be practically certain that the event A will not occur on a single realization of the conditions S.*

Kolmogorov argued that the approach described above, makes these axioms necessary as follows:

Empirical Deduction of the Axioms. Usually one can assume that the system F of events $A, B, C \ldots$ that come into consideration, and are assigned definite probabilities form a field that contains E (Axioms I and II and the first half of Axiom III-the existence of the probabilities). It is further evident that $0 \leqslant m/n \leqslant 1$ always holds, so that the second half of Axiom III appears completely natural. We always have $m = n$ for the event E, so we naturally set $P(E) = 1$ (Axiom IV). Finally, if A and B are mutually incompatible (in other words, the sets A and B are disjoint), then $m = m_1 + m_2$, where m, m_1, and m_2 are the numbers of experiments in which the events $A \cup B$, A, and B happen, respectively. It follows that

$$\frac{m}{n} = \frac{m_1}{n} + \frac{m_2}{n}.$$

So it appears appropriate to set $P(A \cup B) = P(A) + P(B)$.

Remark I. If two assertions are both practically certain, then the assertion that they are simultaneously correct is practically certain, though with a little lower degree of certainty. But if the number of assertions is very large, we cannot draw any conclusion whatsoever about the correctness of their simultaneous assertion from the practical certainty of each of them individually. So it in no way follows from Principle A that m/n will differ only a little from $P(A)$ in every one of a very large number of series of experiments, where each series consists of n experiments.

Remark II. By our axioms, the impossible event (the empty set) has the probability $P(\emptyset) = 0$. But the converse inference, from $P(A) = 0$ to the impossibility of

A, does not by any means follow. By Principle B, the event As having probability zero implies only that it is practically impossible that it will happen on a particular unrepeated realization of the conditions S. This by no means implies that the event A will not appear in the course of a sufficiently long series of experiments. When $P(A) = 0$ and n is very large, we can only say, by Principle A, that the quotient m/n will be very small it might, for example, be equal to $1/n$.

To extend the formalism with conditional probabilities one uses Bayes rule as a definition:

$$P(A|B) = \frac{P(A \cup B)}{P(B)}.$$

This definition now connects to observation in the same way as item 4 by defining the conditions of a system S by additionally requiring that B is observed.

1.1.3 Objective formalism of uncertainty

The objective formalism probabilities are defined relative to a logic scientific model. The induction of probabilities according to this formalism goes back to Laplace [40] and Keynes [33], and is informally characterized by the "Principle of indifference": whenever there is no evidence favoring one possibility over another, they have the same probability. The principle is made more rigorous as follows.

If a scientific model contains a possibly hidden or uncontrolled variable X, and the model does not imply any preference on the values of the variable, one can assign probabilities to these variables as follows:

- if the model implies that X can take maximally n different values x, then $P(x) = 1/n$.
- if the model implies that X can have values in a continuous segment of length l, then $P(a \leqslant x < b) = (b - a)/l$, for a, b in the segment.

The probability of observing an observable value $A = a$ is defined as the sum of all probabilities of X that imply $A = a$. A similar reasoning as in the frequentist case shows now that the Kolmogorov axioms follow from this definition in the finite case. For the infinite case the corresponding arguments are more difficult.

The probability $P(A = a|B = b)$ is defined as the probability $P'(A = a)$ if the rule that observable B has $B = b$ is added to the scientific model. It follows now that

$$P(A = a|B = b) = \frac{P(A = a \cup B = b)}{P(B = b)}.$$

The formalism seems only to allow the introduction of probabilistic rules, for application, one seems to need other formalisms. Under some conditions which we do not want to go into, objective probabilities can be interpreted as item 4 in the previous subsection.

In this section, objective probabilities were defined as uniform distributions. Many logical models imply Gaussian, and other non-uniform models. Suppose that A has a non-uniform distribution P_A, and let $cP_A(a)$ denote the probability of observing $A = a'$ for $a' \leqslant a$, then a hidden variable $B = cP(A)$ defines a uniform distribution. However, one can argue that the definition of such a variable is objective. The advantage of this objective formalism of uncertainty, is that it is very simple.

In conclusion, the frequentist, and objective formalisms lead to a probability that satisfies the same rules, given by Kolmogorov axioms and Bayes axiom. Additionally, under some conditions, they can be applied for prediction of future observable values.

1.2 Statistical hypotheses testing questions

In scientific modeling, one often first starts to infer rules from a restricted context (*inference*), and conjectures that they apply to larger contexts (*generalization*). Scientists agree or disagree on the applicability of rules and models in different contexts. When a rule is under discussion, the rule is called a *hypothesis*. If scientists agree on the applicability of hypotheses and models, science is advancing, and new less probable hypotheses can be discussed.

Statistical hypothesis tests provide a tool that can be useful in the discussion of the applicability of a hypotheses in a restricted context. Such a restricted context is often called an *experiment*, and the values of variable observables is called *data*.

The statistical hypotheses testing questions are:

- *Is the data typical for the model defined by a hypothesis ?*
- *Which of two hypotheses are favored by the data ?*

When frequentist or objective probabilities are involved in a model, a hypothesis can imply a probability distribution over all possible expected observations in a restricted context (data in an experiment). A hypothesis that implies such a probability distribution is called a *simple hypothesis*.

In many actual problems Normal, binomial, and uniform distributions are assumed, where actually a rich structure is available in the data. This happens for example very often in image denoising, while often both the pixels in the noise and the content of the image have much deeper structure than independent Gaussian. Despite this observation, there are interesting applications of such a test, which is often referred as identity testing [55].

The three main contributions of this thesis are now framed within these questions.

- A first approach to the first question is initiated by defining for any data sequence, a set of typical models using Kolmogorov complexity (descriptional complexity). There are close connections with an algorithmic variant of sufficient statistic. These models may not be useful in a direct way to define a procedure for the first question, however, minimal typical models lead to useful technical tools in proofs, and arguments (Chapter 3).

- A second approach of the first question motivates the study of optimal significance tests in some computability class. The existence of such optimal significance tests is easily shown in the case of simple hypotheses. However, related to non-simple hypotheses in many cases an optimal semimeasure is defined. It will be shown that such an optimal semimeasure can in general not be computable. Many interesting questions can be raised on the existence of optimal tests for such semimeasures (Chapter 4).

- The definition of multiplicatively optimal semimeasures lead to a solutions for the second statistical hypotheses testing question (See Chapter 5).

- Coding theorems, and additivity results are provided to characterize tests defined in the previous item, and connect them to data compression heuristics (Chapter 6).

The subsequent Sections 1.3 and 1.4, provide motivation for the studied approaches by relating the answers to statistical literature.

1.3 Is the data typical for a model ?

1.3.1 Typical models

Suppose a hypothesis is formulated with probabilities using the objective uncertainty formalism, then a possibly hidden or uncontrolled discrete variable U, is assumed to be defined. In the discrete case, let $\mathcal{U}$ be a large finite set of possible values for U. The model obtained by the hypothesis is a priori indifferent which $u \in \mathcal{U}$ is preferred. Fix some enumeration of $\mathcal{U}$, and some constant c. With overwhelming objective probability, one can assume that the index of some realization of u is larger than $2^{-c}|\mathcal{U}|$. Let S be the set of all possible observations of an observable A, depending on the possibly hidden or uncontrolled variable U.

Firstly, suppose that there is a bijective connection between $\mathcal{U}$ and S. The enumeration of $\mathcal{U}$ now defines an enumeration of S. Again with overwhelming objective probability the index of x in this enumeration is larger than $2^{-c}|S|$. This motivates the definition of S being a *typical set model* for x iff there is no shorter way within an additive c constant, to algorithmically describe x given S, than by

giving the index of S in some algorithmic enumeration of S. Remark that the length of such a description is larger than $\log |S| - c'$, for some c' large enough.

If there is no bijective connection between between $\mathcal{U}$ and S. Let $P_A(a)$ be the objective probability of $A = a$ according to the hidden or uncontrolled variable U. Now the enumeration of all $b \in S$ is assumed to be in decreasing order of $P_U(b)$. For some c, with overwhelming objective probability the index of x is larger than $2^{-c} P_U(x)^{-1}$. It can be assumed that an algorithmic description of such an index requires $-\log P_U(x) - c$ bits. This motivates to define that P is a *typical probabilistic model* for x iff there is no shorter way to algorithmically describe x given P then $-\log P_U(x) - c$.

1.3.2 Sumtests for simple hypotheses

A hypothesis test d maps each discrete data x to Real numbers $d(x)$, defining an order on the data. The test is constructed such that $d(x)$ is high for data x that seem to contradict the zero hypothesis. A *simple hypotheses* is a hypotheses that implies a unique (objective, or frequentist) probability distribution for all expected data in an experiment. The *significance* of data x according to d is given by:

$$\alpha(x) \quad = \quad \sum \{P^0(y) : d(y) \geqslant d(x)\}$$

If $\alpha(x)$ is small, a scientist will conclude that: *either a rare event has occurred or that P^0 is not representative for x.* In practice, he will reject the zero hypothesis. Therefore $\alpha(x)$ addresses the first hypothesis testing question.

The question rises, whether there is a test T defining an optimal significance $\alpha_T(x)$ in the sense that for any test T' there is a constant c such that for all x: $\alpha_{T'}(x) \leqslant c\alpha_T(x)$. Without loss of generality one can make a monotone mapping of T such that $T(x) = -\log \alpha_T(x)$, which results in the definition of a Martin-Löf test.

A *Martin-Löf test* d for a probability distribution P is a mapping from a discrete set X to $\mathbb{Z}$ such that for any k:

$$\sum \{P(x) : x \in X \wedge d(x) \geqslant k\} \leqslant 2^{-k}.$$

The a prior probability of observing a high value $d(x)$ according to the distribution P for a Martin-Löf test d for P is low. The question rises whether there exist Martin-Löf tests for some P, that is within some computability class maximal within an additive constant. Remark that for such a test, the corresponding significance is multiplicatively optimal.

A function $d : X \to \mathbb{Z}$ is a *sumtest*

$$\sum_{x \in X} P(x) 2^{d(x)} \leqslant 1.$$

In this work the related concept of sumtests will be used. The difference is technical since all conclusions for sumtests, obtained in this manuscript, are also valid for Martin-Löf tests. The use of sumtests increases coherency in the manuscript, and reduces technical details. The question rises whether there exists a sumtest for some P, that is within some computability class maximal within an additive constant.

1.3.3 Sumtests for composite hypotheses

Typically, one is confronted with some family of probability distributions $P(x|\theta)$ depending on θ. A composite hypothesis refers to a hypotheses that corresponds to a set of parameters θ. Here, a slightly different definition of composite hypothesis is used.

A *semimeasure* P is a positive Real function such that

$$\sum P(x) \leqslant 1.$$

It will be argued in Section 2.2 why semimeasures are used, rather then probability distributions. A *composite hypothesis* is a collection of rules that imply a set of semimeasures.

A semimeasure m is *universal* in a set S of semimeasures, if for every $P \in m$, there is a constant c such that $P \leqslant cm$. If the set of semimeasures corresponding to a hypotheses is convex, and countable: $H = P_1, P_2, ...$ for $i = 0, A$, a universal semimeasures m is given by:

$$m =^* \sum_j a_j P_j. \tag{1.1}$$

Since all universal semimeasures are equal up to a constant factor, the subjectivity is limited to such a constant factor.

The "significance" $\alpha(x)$ for some d relative to m, does not have a direct frequentist or objective interpretation. In general, no repetitive experiment with controlled, and uncontrolled variables exists that can result in frequencies given by m. There are also no objective hidden or uncontrolled variables justifying the definition of an objective probability. Any choice of the positive constants a_i in (1.1) results in a universal semimeasure. Suppose that the hypotheses implies a set of computable or lower semicomputable semimeasures that can be enumerated (see further), without loss of generality the constants $a_i = 2^{-K(P_i)}$ can be chosen, with $K(P_i)$ the descriptional complexity of some P_i. Than for $i \leqslant k/(\log k)^2$, and for some large enough fixed constant c, one has

$$P_i^0(x) \leqslant ckm^0(x).$$

If d is large, this means that either:

- the data is generated by some model of high complexity,
- a rare event has occurred,
- the data is not typical for the hypotheses.

In many cases the first interpretation must be partially taken into account, and therefore one should look for a separate notion of significance for the statistic $d(x)$.

1.4 Favoring one of two hypotheses

1.4.1 Subjective belief formalism

For completeness, the subjective belief formalism is given, and the applicability of Kolmogorov's axioms, and the Bayesian axiom discussed.

The subjective belief formalism defines probabilities for one specific person or agent. Such probabilities can directly relate to actions or decisions for the agent. Such probabilities are not directly useful to define probabilistic rules in a scientific model. Since science looks for agreement between many scientists on such rules, the agreement of a single scientist with some rule can be expressed using such a formalism for uncertainty.

The subjective probability $P(A)$ given by an agent to an event A, is the price the agent is willing to pay to play a game against a banker where the agent receives one unit of utility if the event A occurs, and no utility if the event A does not occur. The bet is said to be *fair* if the agent is indifferent between being the player or the banker, once the price has been fixed.

Suppose the following game [21]: a booker offers some bets, the agent decides how much he wants to pay, and subsequently the booker decides whether he is the player or the banker. A *Dutch Book* is a strategy for the booker such that he always wins. It can be observed that a rational agent assigns probabilities such that the Kolmogorov axioms I-IV are satisfied. Also, axiom V must be satisfied such that there is no Dutch Book possible: suppose that for two exclusive events A, B probabilities are assigned such that

$$P(A \cup B) < P(A) + P(B),$$

then the booker can offer bets A, B, $A \cup B$, and subsequently choose to be the banker for the bets of A, B, and be the player for the bet on $A \cup B$. Now the booker always wins.

It can also be shown that the requirement that no Dutch Books exists imply Bayesian rules of conditioning, provided the introduction of possibility of bets being paid back if the conditioned event does not occur, and the Dutch Book argument can be performed in a synchronous way [10, 22].

However, there is some subtle difference between conditional probabilities, and a posteriori probabilities, where a Dutch Book argument must be performed in an asynchronous way. This leads to a deep discussion on Bayesian updating of belief factors, where there is still no agreement [29]. An alternative updating rule has been proposed, which also resists all accepted Dutch Book arguments [60]. However, because of the overwhelming practical successes, and because of simplicity, the Bayesian updating rules will be assumed in the subsequent Subsections 1.4.2 and 1.4.3.

In conclusion, it is observed that the three formalisms of uncertainty result in probabilities satisfying the same Kolmogorov axioms, and Bayesian rule, with some subtle issues for Bayesian belief updating. In many cases, one can identify probabilities obtained by one formalism with probabilities from another formalism, this creates probabilities which have no longer a pure frequentist, objective, or subjective origin, and can be both applied by the frequentist formalism, or by betting strategies.

1.4.2 Favoring one of two simple hypotheses

Let H^0, the *zero hypotheses*, and H^A, the *alternate hypotheses* be two simple hypotheses, implying the frequentist or objective probability distributions P^0 and P^A. A statistical test d is now constructed in order to have high $d(x)$ for data x that seem to favor the alternate hypothesis. Significance, and sensitivity for data x according to d are given by:

$$\begin{aligned} \alpha(x) &= \sum\{P^0(y) : d(y) \geqslant d(x)\} \\ \beta(x) &= \sum\{P^A(y) : d(y) \leqslant d(x)\}. \end{aligned}$$

Remark that α has the same definition as in Section 1.3. If $\alpha(x)$ is small, a scientist will reject the zero hypothesis. If also $\beta(x)$ is small, he will favor the alternate hypothesis. Therefore $\alpha(x), \beta(x)$, jointly address the second hypothesis testing question.

The interpretations of α and β can be given in a compatible way with there frequentist or objective origins. Therefore, for a frequentist setting, the significance respectively sensitivity of a statistical test is the maximal limit fraction of repetitive evaluations of the test where P^0 is rejected, in a context where P^0 describes the observed data well, respectively will not reject P^0 in a context where P^A describes the observed data well. In the objective setting, the significance respectively sensitivity is an objective prior probability that the zero hypothesis disqualifies itself, respectively alternate hypothesis disqualifies itself.

A specific choice of such a statistical test d is given by the likelihood ratio $P^A(x)/P^0(x)$. There are three main motivations to use this test in the case of the testing of two simple hypotheses:

- This ratio also has an interpretation using Bayesian subjective probabilities: if a^0/a^1 represents the ratio of prior belief in the hypothesis corresponding to P^0 relative to the belief in the hypothesis corresponding to P^A, then after observing data x the posterior ratio of the beliefs is:

$$\frac{a^0}{a^1}\frac{P^0(x)}{P^A(x)}.$$

  Remark that observed data Bayesionally updates the beliefs of two scientists.

- Due to the Coding Theorem [44] (see further) the previous Bayesian interpretation is also justified by an Occam's razor argument that favors the hypothesis that can be described with minimal code length.

- The Newman-Pearson lemma states that $\beta \circ \alpha^{-1} : [0,1] \rightarrow [0,1]$ is uniformly maximal for

$$d(x) = P^A(x)/P^0(x).$$

  This means that there is a test that has for any significance an optimal sensitivity.

This shows that optimal hypothesis testing is equivalent to likelihood ratio testing. Remark that significance, and sensitivity are bounded by $P^0(x)/P^A(x)$.

1.4.3 Favoring one of two composite hypotheses

There is no accepted optimal general way to determine a preference for one of two composite hypotheses tests. Many methods in literature are proposed that are theoretical optimal under some conditions [19], or have been found to be empirically useful in specific contexts. Let H^0 and H^A be the sets of semimeasures constituted by the zero and alternate hypothesis.

- *Uniformly optimal test:* In specific cases, there is a test that has an optimal $\beta \circ \alpha^{-1}$ function for all combination of tests in H^0 and H^A. Remark that this is the case for many important hypotheses tests for normal distributions.

- *Bayesian approaches*: Assign some fixed prior probability to all semimeasures, this reduces the problem to simple hypothesis testing. Often it is not possible to extend the hypothesis with an acceptable prior, and therefore this is a subjective method.

- *Generalized maximal likelihood*: This is the likelihood ratio of the best case hypotheses:

$$\frac{\max\{P(x) : P \in H^A\}}{\max\{P(x) : P \in H^0\}}.$$

This is the most commonly used method, and can also be applied to obtain some important tests for normal distributions. In specific cases this method is proved to be optimal, but in other cases it has problems or is subject to discussion [19].

Suppose that a composite zero H^0 and alternate hypothesis H^A have universal semimeasures m^0 and m^1. If the sets of the semimeasures are convex and countable $H^i = P_1^i, P_2^i, ...$ for $i = 0, A$, as mentioned earlier, the universal semimeasures m^i satisfies:

$$m^i =^* \sum_j a_j P_j^i.$$

This means that hypothesis selection by the likelihood ratio

$$d(x) = m^A(x)/m^0(x),$$

can be considered as a Bayesian approach to composite hypothesis selection. Since all universal semimeasures are equal up to a constant factor, the subjectivity is limited to a constant factor. (Thus if the scientist supporting the zero or alternate hypotheses, they can maximally disagree within a constant Bayesian factor.) Also, because $m^i \in H^i$, and because it is multiplicatively optimal, it can be considered as generalized maximal likelihood testing if one can neglect the constant factors.

Again the significance of d relative to m^0, does not have a direct frequentist or objective interpretation. Assume H^0 and H^A satisfy the conditions of Equation (1.1). Suppose that the hypotheses implies a set of computable or lower semicomputable semimeasures that can be enumerated (see further). Since all universal semimeasures are equal within a constant additive multiplicative factor, without loss of generality the constants $a_i = 2^{-K(P_i)}$ can be chosen, with $K(P_i)$ the descriptional complexity of some P_i. Then, for $i \leqslant k/(\log k)^2$, and for some large enough fixed constant c, one has

$$P_i^0(x) \leqslant ckm^0(x).$$

Since any choice of the positive constants a_i in (1.1) results in a universal semimeasure, without loss of generality the constants $a_i = 1/i(\log i)^2$ can be chosen. Let $i \leqslant k/(\log k)^3$ then:

$$\frac{P_i^0(x)}{m^A(x)} \leqslant^* \frac{km^0(x)}{m^A(x)}.$$

If the significance of d is large, this means that either:

- Some complex model from the zero hypothesis describes the data.
- The alternate hypothesis m^A more adequately describes the data.
- A rare event has occurred.

In many cases the first interpretation must be partly taken into account, and therefore one should look for a separate notion of significance for the statistic $d(x)$. For example, a frequentist significance bound is obtained by a permutation test for the Shannon information transfer statistic in [52].

Conclusions

The goals and procedures in scientific modeling have been outlined, as a search for scientists to find agreement in logical rules for reproducing past observations, and controlling future observations. Within these goals two formalisms of uncertainty are used: frequentist and objective. The corresponding probabilistic hypotheses lead to two questions addressed in statistical hypotheses testing: "Is the observed data typical for a model ?", and "Which of two hypothesis favors the data ?"

For the first question two answers are given: by the definition of typical models, and by the definition of sumtests. Typical models have a direct interpretation in terms of the objective uncertainty formalism, while sumtests relate to the common use of significance tests. Universal semimeasures are introduced, which takes over the same procedures of the simple approach with a slightly different interpretation of the significance.

To address the question of favoring one of two hypotheses, first the formalism of subjective probabilities is discussed. The goal of such tests is to convince other scientists to lower or increase there relative subjective beliefs in one of two hypothesis for a context. Subjective beliefs may be influenced by the observation of data in an experiment. For simple hypotheses there seems to be a clear logic in how to do this by the use of ratio tests. For composite hypotheses, such an approach is not available, however on the conceptual level a solution is given by ratio tests of universal semimeasures corresponding to the hypotheses.

2

Computability and Kolmogorov complexity

Abstract. The section introduces computability theory and Kolmogorov complexity, and also provides some straightforward extensions of the theory for later use.

Two equivalent formalisms of computability are given, the Turing computable formalism, and the more mathematical recursive functions formalism. Turing computability appeals directly to intuition while recursive functions define a very powerful class of indexed universal interpreters, which allows to efficiently handle some concept of "limited computation".

Lower semicomputable semimeasures are defined, and is shown that they provide a unique way to define universal semimeasures corresponding to most hypotheses. Finally Kolmogorov complexity for a discrete finite sequence is defined as an ideal data compression length of a discrete finite sequence. It satisfies an additivity property, and a Coding Theorem, which are the two cornerstones of algorithmic information theory.

2.1 Computability I

Church's thesis can be stated that any deterministic physical device, that can be considered to map discrete input into discrete output, has the same symbolic input-output mapping as some Turing machine. Church thesis, allows to objectively answer the question: "What can be computed ?". Moreover, all reasonable mathematical formalizations of "computability" lead to equivalent definitions. Two such

formalizations are partial recursive functions, discussed in Section 2.3, and Turing partial computable functions, discussed in this section. The first allows to observe that mathematically defined functions which do not use $\forall$ and $\exists$ symbols, can be computed, while the latter allows the use of our intuition obtained by using and programming modern laptops and PC's, to understand whether specific algorithms define computable functions.

2.1.1 Notation

A good introduction to computability theory is [67]. More background can be found in the textbooks [50, 51, 62], from which our notation is obtained. ω is the set of natural numbers $\{0, 1, 2, \dots\}$. $\omega^{<\omega}$ is the set of finite sequences of natural numbers, denoted as $[x_1, \dots, x_n]$, and also as $x_1 \dots x_n$. Functions $\omega \times \cdots \times \omega$ are often identified with functions over $\omega^{<\omega}$. Concatenation of sequences $a, b \in \omega^{<\omega}$ is denoted as ab.

A partial function over a set S is a function that is defined on a subset of S. A total function over S, is a function that is defined on all elements of S. If the value $f(x)$ of a partial function f is defined, then it is written $f(x)\downarrow$, otherwise $f(x)\uparrow$.

Remark that all mathematical results obtained from the literature are denoted by "Theorem", while non-trivial results introduced here, or in one of the author's papers, are referred to as "Propositions". Results that explore simple properties of definitions, or technical results for later use in proofs, are referred as "Lemmas".

2.1.2 Partial computable functions

A *Turing machine* consists of a bidirectional infinite tape, and a program, carrying out some computations. The description of a Turing machine in [67] is cited here with minor customizations.

Formally, the instructions in the program of a Turing machine are quintuples

$$(q_i, x, q_j, y, X),$$

where q_i and q_j are from a finite set Q of states*, x and y are either 0 or 1, and $X \in \{L, R\}$. The interpretation of such an instruction is that if the machine is in a state q_i and scanning a cell of the tape with symbol x, then the contents of the cell is changed to y, the next machine state is q_j and the tape head moves one step in the direction X, where L stands for "left" and R for "right". A Turing machine program is a set of instructions saying what to do in any of the possible situation the machine may be in. It is therefore defined by a function*

$$M : Q \times \{0, 1\} \to Q \times \{0, 1\} \times \{L, R\}.$$

Furthermore, the set Q contains two distinguished states: an initial state q_0 and a final state q_f. The interpretation is that the machine starts a "computation"

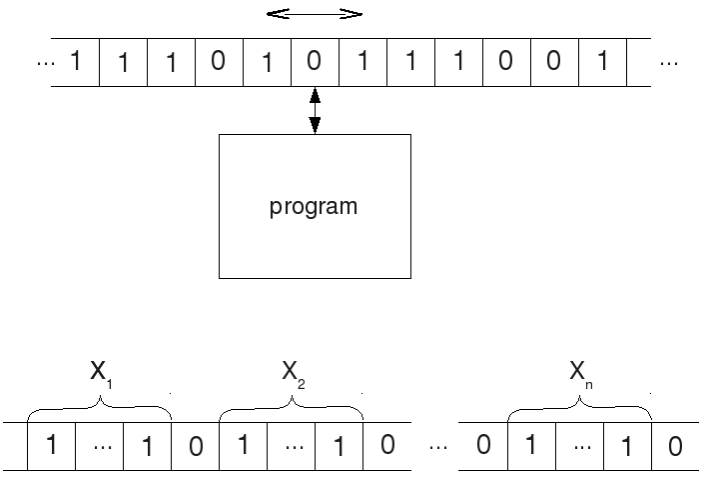


Figure 2.1: Turing Machine

in the state q_0 and halts when it reaches the final state q_f. We can think of Turing machines as idealized computers, with unlimited resources such as time and memory.

A function $f : \omega^{<\omega} \to \omega^{<\omega}$ is partial computable if there is a Turing machine such that for every $x \in \omega^{<\omega}$ with $f(x) \downarrow$, one has that when x is written on tape, as in figure 2.1, and a computation is started with the program above a bit of x_1, then the program reaches the Halting state while the sequence $f(x)$ is written on the output tape, also as in figure 2.1. In this way, a Turing machine can be associated with its partial function. A *partial computable function* is a function associated with a Turing machine in such a way. The expression

$$\phi_t(x) \downarrow = y \tag{2.1}$$

means that a Turing machine ϕ reaches the state q_f after t computation steps.

Valid programs in almost any programming languages such as $C++$, $python$, ... can be considered as terminating programs on a Turing machine. Remark that here the "program" of a Turing machine is actually, the whole hardware of the computer combined with the software of the operating system and a compiler for the programming languages. In the same way as for $C++$, $python$, ..., on such machines on can store and load large numbers, do arithmetic, use for and while loops, call functions, ... This provides a general technique to proof that functions are computable. Nowadays, the programming analogy provides a strong intuition for what universal Turing machines can do.

It can be shown that there exists a Turing machine ϕ, that for other Turing machine ψ and for any x can simulate the execution of $\psi(x_1, \ldots, x_n)$, by exe-

cuting $\phi(e, x_1, \ldots, x_n)$. More formal, for any ψ, there is an e such that for all $x \in \omega^{<\omega}$ one has $\phi(ex) = \psi(x)$ if $\psi(x) \downarrow$. For the contemporary reader this is not surprising, since modern computers can simulate almost any deterministic physical process, can run virtual machines and almost any computer languages can be compiled into each other. However, it is still remarkable that such machines can be very simple. Using the formalism of "combinators" such a machine can be defined within 54 bytes [44] !

2.1.3 Semicomputable and Real functions

A function f into ω^2 can be considered as a function into $\mathbb{Z}$ as follows: $f(x) = [r, s] = (-1)^r s$. A function f into $\mathbb{Z} \times \omega \backslash \{0\}$ can be interpreted as a function into $\mathbb{Q}$ as follows: $f(x) = [r, s] = r/s$.

The computable functions into $\omega^{<\infty}$, $\mathbb{Z}^{<\infty}$, or $\mathbb{Q}^{<\infty}$, are denoted by Δ_1. A two-argument Δ_1-function $g_t(x)$ is an *approximation* of a function $f(x)$ if $\lim_{t\to\infty} g_t(x) = f(x)$. The two-argument function $g_t(x)$ is an *enumeration* of f iff g is an approximation of f, and for any t: $g_t(x) \leqslant g_{t+1}(x)$. $g_t(x)$ defines a *co-enumeration* for f if $-g_t(x)$ is an enumeration for $-f$. The set of approximations, respectively, enumerations, and co-enumerations is denoted with $\dot{\Delta}_2$, respectively, $\dot{\Sigma}$, and $\dot{\Pi}$. The set of functions that have an approximation, respectively, enumeration, and co-enumeration are called limit-computable, respectively, lower semicomputable, and upper semicomputable functions, and are denoted with Δ_2, respectively, Σ, and Π. It can be proved that also for all these function types one has $\Delta_1 = \Pi \cap \Sigma$[1].

While the manuscript could have been written using only Rational functions in stead of Real functions, for the purpose of uniformity with the literature Real functions will be used.

As common in computability theory, the Real numbers in the interval $[0, 1]$ are associated with Cantor space 2^∞, which are the infinite binary sequences. Remark that this association is not bijective since for example for Real numbers one has $0.0111\cdots = 0.1000\ldots$ however, in Cantor space this equality does not hold. Also, for this work, this omission does not create any problems.

Remark that also Real functions can have approximations, enumerations, and co-enumerations. The classes of Real Δ_2, Σ, and Π-functions remains the same as the definition of the corresponding rational function classes. The class of Real Δ_1-functions is defined as $\Delta_1 = \Pi \cap \Sigma$. Remark that the class of one argument Δ_1-functions f into $\mathbb{R}$ is exactly the class of functions for which there is a two argument Δ_1-function g into $\mathbb{Q}$ such that for any x and $k \in \omega$: $\text{abs}(f(x) - g(x, k)) \leqslant 2^{-k}$.

[1] Remark that this is a rather complicated for Rational functions.

It can be shown that for all function types one has for $i = 1, 2$, $\Delta_i \neq \Sigma$, and $\Delta_i \neq \Pi$ [67].

2.2 Semimeasures

It is shown that the classes of Σ-semimeasures corresponding to a large group of hypotheses, have universal elements, while either the classes of Δ_1, Σ, and Π-measures, or the classes of Δ_1 and Π-semimeasures have no or only in special cases universal elements.

A (discrete) *semimeasure* $P : \omega \to [0, 1]$ is an Σ-function that satisfies

$$\sum\{P(x) : x \in \omega\} \leqslant 1.$$

2.2.1 Computability classes of semimeasures

The classes Δ_1, Σ, and Π-measures, reduce to two strictly different classes.

Proposition 2.2.1 ([7])**.**

1. *Every Σ-measure is computable,*
2. *There is a Π-measure that is not computable.*

Proof.
1. This is well-known: If $P \in \Sigma$ with computable approximation P_s and $\sum_x P(x) = 1$ then to approximate $P(x)$ to within ε, find a stage s such that $1 - \sum_x P_s(x) < \varepsilon$. Then $P(x) - P_s(x) < \varepsilon$.

2. A family of sets X_s is computable if there is a two argument Δ_1-function f such that f_s is the characteristic sequence of the set X_s. A Π set X is a set for which there is a computable family of sets X_s such that $X_0 \supset X_1 \supset \ldots$ and $X = \Cap_s X_s$. Let X be any non computable Π-set, with such computable approximation X_s. Define a measure P as follows: At stage s assign the s values $2^{-1}, \ldots 2^{-s}$ to the first s elements of $X_s \subseteq X_{s-1}$, in such a way that the elements of X_s that were already assigned a value at a previous stage retain this, and the values that were assigned to elements in $X_{s-1} - X_s$ are given a new host element. For any element $x \notin X$ we define $P(x) = 0$. Then $P \in \Pi$, and P is not computable because otherwise, since $x \in X \Leftrightarrow P(x) > 0$, X would also be computable. Note that in general $P(x) > 0$ is not decidable for computable P, but in this case it is: x is assigned an initial value 2^{-i} with $i \leqslant x$. Computing $P(x)$ to within precision 2^{-i-1} decides whether it is 2^{-i} or 0. □

Remark that by this proposition, also the class of Δ_1-semimeasures differs from the classes of Π-semimeasures. The Corollary 2.2.3 the class of Δ_1-semimeasures

differs from the Σ-semimeasures. This shows that there are three different classes of semicomputable semimeasures: Δ_1, Σ, and Π-semimeasures.

2.2.2 Classes with universal semimeasures

A semimeasure P (multiplicatively) *dominates* a semimeasure Q iff:

$$\exists c \forall x [cP(x) \geqslant Q(x)].$$

For S a set of semimeasures, a semimeasure m is *universal* in S iff m dominates all semimeasures in S. By Proposition 2.2.2, the Δ_1 and Π-semimeasures, and all measures do not define universal semimeasures.

Proposition 2.2.2 ([7]). (i) *There is no universal Δ_1-measure and Δ_1-semimeasure.*

(ii) *There is no universal Π-measure and Π-semimeasure.*

Proof. Both item (i)-(ii) follow from the following. Let P be a Π-semimeasure. We construct a computable semimeasure Q such that

$$\forall q \in \mathbb{Q}^{>0} \exists x \;\; P(x) < qQ(x). \tag{2.2}$$

Given q we simply search for an x where $P(x)$ is small, and set a large value for $Q(x)$. Note that x can be found effectively since $P \in \Pi$. More precisely, given $q = 2^{-i}$ find a fresh x such that $P(x) < 2^{-2i}$. Set $Q(x) = 2^{-i}$, and to make Q total set $Q(y) = 0$ for all $y < x$ that were not yet defined. The Q thus constructed is computable, clearly satisfies (2.2), and $\sum_x Q(x) = \sum_i 2^{-i} = 1$. □

Corollary 2.2.3 ([7]). *Let m be the universal Σ-semimeasure, and let P be a Π-semimeasure. Then the function $m(x)/P(x)$ is unbounded.*

Proof. Suppose for a contradiction that $c \in \omega$ is a constant such that $m(x)/P(x) \leqslant c$ for every x. Then $m(x) \leqslant cP(x)$, and P dominates all Σ-semimeasures. Consequently it dominates all Δ_1-semimeasures, and therefore it is a universal Δ_1-semimeasure, contradicting Proposition 2.2.2. □

In contrast with Δ_1 and Π-semimeasures, the Σ-semimeasures define a universal element.

Theorem 2.2.4. *There exists a universal Σ-semimeasure.*

Proof. Let for every $i \in \omega$ the semimeasures P_i be defined by

$$P_{i,t}(x) = \begin{cases} 0 & \text{if } t = 0 \\ \max\{P_{i,t-1}, \phi_t(i|x)\} & \text{if } \sum\{P_t(x) : x \leqslant t\} \leqslant 1 \\ P_{i,t-1} & \text{otherwise.} \end{cases}$$

Remark that P_i defines a Σ-semimeasure. Let Q be a Σ-semimeasure such that $Q(i) > 0$ for all i, and let

$$Q_U(x) = \sum_i Q(i)P_i(x). \tag{2.3}$$

Q_U defines a Σ-semimeasure, and dominates all Σ-semimeasures. □

2.2.3 Ideal sequence prediction

For any $x \in 2^\infty$, the notation $x^i = x_1 \dots x_i$ is used. Let $x^i b$ means appending a bit $b \in \{0,1\}$ to the binary string x^i. x is computable iff the function $f : i \to x_i$ is computable. For $x \in 2^\infty$ the task of sequence prediction is the task of predicting x_{i+1}, given x^i for all $i \in \omega$. Suppose that for a universal semimeasure m, one predicts $\hat{x}_{i+1} = 1$ if

$$\frac{m(x^i 1)}{m(x^i 0)} \geqslant 1,$$

and $\hat{x}_{i+1} = 0$ otherwise. For a computable sequence x the amount of errors according to the prediction strategy $\hat{x_i}$ is bounded by the length of a program for the function $f : i \to x_i$ [44].

If x is generated stochastically, according to some computable P, then the expected amount of deviation of the Bayesian conditionals of $m(x_{i+1}|x^i)$, and $P(x_{i+1}|x^i)$ converge rapidly to each other for increasing i. More formal, the expectancy according to P of $\sum_i (P(x_{i+1}|x^i) - m(x_{i+1}|x^i))^2$ is bounded by $c \ln \sqrt{2}$, where c is the constant such that $P \leqslant cm$ [63]. This shows that m is expected to converge very rapidly to P.

For sequence predicting according to a universal semimeasure with arbitrary loss functions, it was shown that under weak conditions predictors based on m performed only worse by an additive square root term of a normalized reward [30,31]. Since sequence prediction by arbitrary loss functions is mathematically equivalent with problems such as pattern recognition and reinforcement learning, this formally defines optimal solutions for these problems. There are still several interesting open questions relating universal semimeasures, and game theoretic interpretations of algorithmic probability. One such question is whether the convergence in the previous paragraph also appears for individual stochastic sequences, instead off convergence on the average [31, 32].

2.2.4 Universal Σ-semimeasures for a hypotheses

Similar conclusions of the previous subsections also hold for sets of semimeasures corresponding to some large group of non-simple hypotheses. Here the result of Theorem 2.2.4 is generalized.

Definition 2.2.5. Let S be a set of semimeasures:

- $S^{\uparrow}$ is the subset of Σ-semimeasures in S.
- S is *testable* iff there is a computable logic expression L such that for any semimeasure P: $P \in S$ iff some rational approximation P_t of P satisfies:

$$\forall t \forall n \leqslant t\Big[L(P_t^n)\Big],$$

  where P_t^n is the finite restriction of P_t on 2^n.
- S is *convex* iff from any $P, Q \in S$, and $a, b \in [0,1]$ with $a + b \leqslant 1$: $aP + bQ \in S$.
- S is enumerable if a series of codes for $S = P_1, P_2, \ldots$ can be enumerated.
- The product set of two sets of semimeasures S, T is given by

$$S \times T = \{PQ : P \in S \wedge Q \in T\}$$

Remark that the product of two semimeasures is not necessarily a semimeasure.

Proposition 2.2.6. *Let S, T be sets of semimeasures.*

(i) If S is testable and contains $P_0 = 0$ then $S^{\uparrow}$ enumerable.

(ii) If S is convex and $S^{\uparrow}$ can be enumerated as $P_1, P_2, ...$, then $S^{\uparrow}$ contains the universal semimeasure

$$m^{S^{\uparrow}} = \sum a_i P_i,$$

where $a_i > 0$ is any computable real function such that $\sum_{i \in \omega} a_i \leqslant 1$.

(iii) If products of semimeasures in the sets S and T define semimeasures, and if $S^{\uparrow}, T^{\uparrow}$ have universal elements m^S, m^T, then

$$m^{S^{\uparrow} \times T^{\uparrow}} = m^S m^T$$

is a universal element for $S^{\uparrow} \times T^{\uparrow}$.

Proof. The first two items of the proposition are a direct generalization of the proof of the existence of universal Σ-semimeasures [25, 43, 44].
Part (i). Define the enumeration $P_{i,t}$: let $P_{i,0}(x) = 0$ for all i, x. Remark that $P_{i,0} \in S$. For all t let

$$P_{i,t}(x) = \max\{\phi_t(i, x, s) : s \leqslant t \wedge \phi_t(i, x, s) \downarrow\},$$

if $\sum\{P_{i,t}(x) : x \in 2^n\} \leqslant 1$, and $L(P_{i,t}^n)$ is true, otherwise let for all x:

$$P_{i,t}(x) = P_{i,t-1}(x).$$

Remark that for all i, t, $P_{i,t}(x)$ is computable, and that if i is a code for a $Q \in S^{\uparrow}$, then there is an i such that $Q = P_i$.

Part (ii). Let:

$$m_t^S = \sum\{a_i P_{i,t} : i \leqslant t\}.$$

Remark that m_t^S is computable, that it increases with t, and therefore m^S is a Σ-semimeasure. Remark that by convexity for all t, $m_t^S \in S^{\uparrow}$, and for any n, the values $m_t^S(w)$ with $w \in 2^{<n}$ remain constant for some t large enough. Therefore the limit is also in $S^{\uparrow}$. Finally remark that m^S dominates all P_i.

Part (iii). Clearly $m^{S^{\uparrow} \times T^{\uparrow}} \in S^{\uparrow} \times T^{\uparrow}$. Let $R \in S^{\uparrow} \times T^{\uparrow}$. It remains to show that $R \leqslant^* m^{S^{\uparrow} \times T^{\uparrow}}$. There exist $P \in S^{\uparrow}, Q \in T^{\uparrow}$ such that $R = PQ$. Since $P \leqslant c_P m^S$ and $Q \leqslant c_Q m^T$, we have that

$$R = PQ \leqslant c_P c_Q m^S m^T = c_P c_Q m^{S^{\uparrow} \times T^{\uparrow}}.$$

□

From Proposition 2.2.6 it follows that the set of univariate, bivariate and conditional Σ-semimeasures have a universal element denoted as: $m(x)$, $m(x, y)$, and $m(x|y)$. The set of independent Σ-semimeasures is given by $P(x, y) = Q(x)R(y)$, for Q, R univariate semimeasures. The set satisfies the conditions of item (iii) of Proposition 2.2.6, and therefore has universal element $m(x)m(y)$. Also, remark that by Corollary 5.1.5 there are sets S, T, such that the universal element of $S^{\uparrow} \times T^{\uparrow}$ can be a factor $o(n/\log n)$ smaller than the universal element of $(S \times T)^{\uparrow}$.

2.3 Computability II

Computability theory is defined by an older formalism, which will define indexed interpreters. This formalism allows to define some alternative for time-bound Kolmogorov complexity which allows to reduce many technical details.

2.3.1 Partial recursive functions

The class of primitive recursive functions is defined, and it is mentioned that it equals the classes of computable functions. In analogy with partial computable functions, there are universal partial recursive functions.

Definition 2.3.1. The class of *primitive recursive functions* [37] is the minimal class of functions from $\omega^{<\omega}$ into $\omega^{<\omega}$

1. containing the initial functions

$$\begin{array}{rcll} O : x & \rightarrow & 0 & \text{constant zero function} \\ S : x & \rightarrow & x+1 & \text{successor function} \\ \pi_n^i : (x_1, \ldots, x_n) & \rightarrow & x_i & \forall n \forall i \leqslant n \text{ projection function,} \end{array}$$

2. closed under the schemes of composition and concatenation

$$\begin{aligned} f(x) &= h(g_1(x), \ldots, g_n(x)) \\ f(x) &= [g_1(x), \ldots, g_n(x)], \end{aligned}$$

3. closed under primitive recursion

$$f(x, n+1) = h(f(x, n), x, n). \tag{2.4}$$

All primitive recursive functions are total functions. Not all functions that are intuitively considered to be generated by an "algorithm" correspond to total functions. A larger class of partial functions is now defined. For any $y \in \omega^{<\omega}$, and any binary expression R, let $(\mu x)\Big[R(x, y)\Big]$ denote the least x such that $R(x, y) = 1$.

Definition 2.3.2. The class of *partial recursive functions* is the minimal class of functions that

1. satisfies the conditions of the primitive recursive functions,
2. is closed under μ-recursion

$$\phi(x) = (\mu y)\Big[(\forall z \leqslant y[f(x, z) \downarrow]) \wedge f(x, y) = 0\Big].$$

The *recursive functions* are defined by the partial recursive functions that are total in $\omega^{<\omega}$. Remark that primitive recursive functions are computable functions.

Theorem 2.3.3 (Enumeration Theorem)**.** *There exists a partial recursive function ϕ such that for any partial recursive function f there is an e such that*

$$\phi([e, x_1, \ldots, x_n]) \downarrow = f([x_1, \ldots, x_n]).$$

A ϕ satisfying this theorem is called a universal partial recursive function. It can be shown that the class of partial computable and partial recursive functions are exactly the same [68]. This also implies that the class of computable and recursive functions are identical. This is the cornerstone of computability theory.

2.3.2 Prefix-free indexed interpreters

In this work, some notion of Kolmogorov complexity (see further) with a "computation time" restriction will be needed in many technical Lemma's.[2] Traditionally, for such purposes the s-step finite approximation, of partial recursive functions, or a time restriction on a Turing machine is used. However, using indexed interpreters, one is able to save in notation. To illustrate such a simplification, let ϕ be a Turing machine with corresponding partial computable function ϕ as defined in equation (2.1). Let

$$K_t(x) = \min \{l(p) : \phi_t(p) \downarrow = x\} .$$

With full logical notation, the equivalent statement of Proposition 2.4.5, using the traditional version of time bounded additivity of Kolmogorov complexity is

$$\begin{aligned} &\exists c \in \omega \exists f, g \in \Delta_1 \forall x, y, t \in \omega \\ &\left[K_t(x,y) + c \geqslant K_{f(t,x)}(x) + K_{g(t,x,y)}\left(y|x, K_{f(t,x)}(x)\right)\right] . \end{aligned}$$

Using the formulation of indexed interpreters, this is simplified as: "For suitable ϕ

$$\begin{aligned} &\exists c \in \omega \forall x, y \in \omega \forall s \geqslant \log y \\ &[K_{s-1}(x,y) + c \geqslant K_s(x) + K_s(y|x, K_s(x))]\text{"}, \end{aligned}$$

which is much less cumbersome. Especially in Chapter 4, such accumulated simplifications will be substantial. Also, remark that the formulations of many results are independent from the choice of reference machine, therefore in the proofs a "suitable ϕ" can always be assumed.

Some more notation is now given. The finite binary strings are referred as $2^{<\omega}$. The empty sequence is denoted as ε, and the natural bijection between $\omega \leftrightarrow 2^{<\omega}$:

$$0 \leftrightarrow \varepsilon, 1 \leftrightarrow 0, 2 \leftrightarrow 1, 3 \leftrightarrow 00, 4 \leftrightarrow 01, \ldots$$

is assumed. Remark that also natural bijections exist between $\omega^{<\omega} \leftrightarrow \omega$, and $\omega^n \leftrightarrow \omega$ for any n. Concatenation of bits $x_i, i < n$ is denoted as $x_1 x_2 \cdots x_n \in 2^n$. Also, the concatenation of $x \in S^m$ and $y \in S^n$ is denoted as $xy \in S^{m+n}$. The interpreter from [25] is extended with an index.

Definition 2.3.4. An *indexed interpreter* ϕ, is a primitive recursive function into $\omega^{<\omega} \cup \{\infty\}$:

$$\phi : \omega \times 2^{<\omega} \times \omega^{<\omega} \to \omega^{<\omega} \cup \{\infty\} : (t, p, x) \to \phi_t(p|x),$$

such that for any t, p, x and $x \neq \infty$: $\phi_t(p|y) = x$ implies $\phi_{t+1}(p|y) = x$.

[2] For the reader familiar with time bound Kolmogorov complexity, it is remarked that an time abstract index is used, similar to [46].

The formula $\phi_t(p|x) \downarrow = y$ means that $\phi_t(p|x) = y$ and $y \neq \infty$. For ε, the shorthand notation $\phi_t(p) = \phi_t(p|\varepsilon)$ and $\phi(p|x) = \lim_{t\to\infty} \phi_t(p|x)$ are assumed. Every indexed interpreter defines a partial recursive function into $\omega^{<\omega}$. Using the Kleene Normal Form [37, 50, 67], it can be easily observed that every partial recursive function defines an indexed interpreter. Let

$$t_\phi[p|x] = \begin{cases} \min\{t : \phi_t(p|x) \downarrow\} & \text{if } \phi_t(p|x) \downarrow \\ \infty & \text{otherwise.} \end{cases}$$

The index t of an indexed interpreter ϕ_t can be informally considered as a time parameter by Lemma 2.3.5.

Lemma 2.3.5. *For any indexed interpreter ϕ, and for any universal Turing machine ψ there is a computable function f such that*

$$\phi_t(p|x) \downarrow = \psi_{f(t,p,x)}(p|x),$$

if $\phi_t(p|x) \downarrow$.

Proof. ϕ is defined by some primitive recursive function into $\omega^{<\omega} \cup \{\infty\}$. Therefore, there is a program q on ψ that simulates ϕ. Let $f(t, p, x) = t_\psi[q|t, p, x]$, which is always defined since ϕ represents a computable function. □

From now on indexed interpreters are used for formal definitions.

- An indexed interpreter ϕ is *prefix-free* if for any $y \in \omega^{<\omega}$, the set H_y of p such that $\phi(p|y) \downarrow$ is prefix-free.
- A prefix-free indexed interpreter ϕ is *optimal universal*, if there is a $w \in 2^{<\omega}$ such that for all prefix-free indexed interpreters ψ and for all p, w: $\phi(wp|y) = \psi(p|y)$.

From now on an *o*ptimal *u*niversal *p*refix-free *i*ndexed interpreter is referred as an *oupi*-interpreter.
A *free command string* of an oupi-interpreter ϕ is a $w \in 2^{<\omega}$ such that for any p, x: $\phi(wp|x) = \infty$. It will be assumed that any oupi-interpreter has an infinite amount of free commands available. ϕ can be extended with an assignment of a free command w. to define a new oupi-interpreter ψ such that $\psi_t(p|x) = \phi_t(p|x)$ when p does not start with w and such that $\psi_t(wp|x)$ is defined as a primitive recursive function of ψ into $\omega^{<\omega} \cup \{\infty\}$.

A *well-defined command* is a primitive recursive function of ψ where an double induction scheme is used: first induction on s then induction on p. This means that an inductive function defining $\psi_t(wp|x)$ is a well-defined command iff the inductive function only depends on t, p, x, on $\psi_s(q|y)$, for $s < t$, and on $\psi_s(q|y)$ for $s = t$ and $q < wp$. The following well-defined commands assigned to some free command strings w, are assumed at some places in the paper.

- *Measurement of "computation" time.*

$$\phi_s(wp|x) = \begin{cases} t[p|x] & \text{if } \phi_s(p|x) \downarrow \\ \infty & \text{otherwise.} \end{cases}$$

  Remark that $t[wp|x] \leqslant \phi(wp|x)$, when defined.

- *Iterated function call.*

$$\phi_s(wpE(r)|i) = \begin{cases} \phi_s(p|\phi_s(wpE(r-1)|i)) & \text{if } r \geqslant 0, \\ i & \text{otherwise.} \end{cases}$$

- *The lexicographic first string in 2^l, incompressible in time $s-1$.*

$$x = \phi_s(w|s,l) = \min\{x \in 2^l : \forall p \in 2^{<l}[\phi_{s-1}(p|l) \neq x]\}.$$

The expression "For suitable ϕ: EXPR" means that well-defined commands exists such that when added to oupi-interpreter ϕ, the expression EXPR is true. Remark that a more flexible definition of oupi-interpreter could have been chosen, such that the statement "For suitable ϕ" could be removed. However, to remind the the reader one has used the additional freedom of defining recursive well-defined commands, the explicit notation "For suitable ϕ" is used.

2.3.3 Prefix-free Turing machines

A variant of a Turing machine can be defined that is useful for informal interpretation of online Kolmogorov complexities in Chapter 6.

Let ϕ be a Turing machine as defined above. Assume now that ϕ has additionally to its work tape an extra program input tape, which is one-sided infinite. The program can not write to this tape, only read, and the tape can only move forward as in Figure 2.2. $\phi(p|x) \downarrow= y$ iff the computation of p started while the program reeds the first bit of p on the program input tape, and the first bit of x_1 on the work tape, then the program of ϕ attains the Halting state after exactly all bits of p have been red from the input tape, and the work tape shows y. Remark that such a ϕ has for every x a prefix-free set of Halting programs.

2.4 Kolmogorov complexity

For references and background on Kolmogorov complexity it is referred to the excellent textbook [44], and the online notes [25]. From now on a *fixed* oupi-interpreter ϕ will be assumed.

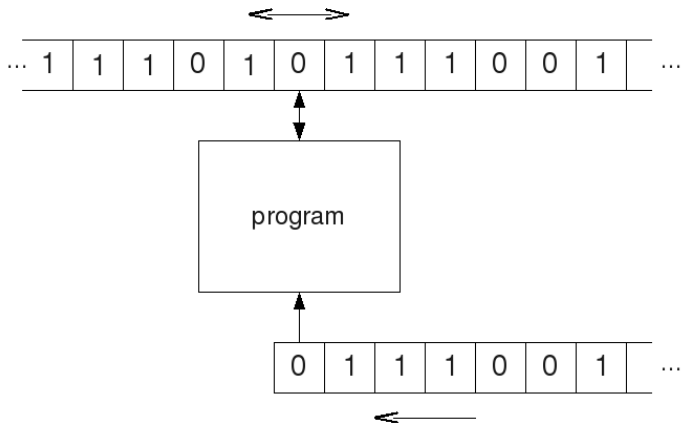


Figure 2.2: Prefix-free Turing machine

Definition 2.4.1. For $x, y \in \omega^{<\omega}$, the *(indexed conditional prefix-free) Kolmogorov complexity* of x given y is

$$K_t(x|y) = \min\{l(p) : \phi_t(p|y) = x\}.$$

If the index t is omitted, the limit t to infinity is assumed. The Kolmogorov complexity of a computable function f is

$$K(f) = \min\{l(p) : \forall x \big[\phi(p|x) \downarrow = f(x)\big]\}.$$

For $F \in \{\Pi, \Sigma, \Delta_2\}$, the Kolmogorov complexity of a function $f \in F$ is given by the minimal Kolmogorov complexity $K(g)$ of a function $g \in \dot{F}$, defining an approximation for f.

Remark that K defines a Π-function. When the requirement of ϕ being prefix-free is dropped, the resulting complexity is called plain Kolmogorov complexity. This complexity will be used in chapter 3. Kolmogorov complexity is a very useful combinatorial tool in the study of statistics and computability. Its use is due to its additivity property, and the Coding Theorem.

2.4.1 Elementary properties of Kolmogorov complexity

For functions f, g the shorthand notation $f \leqslant_+ g$, and $f(x) \leqslant_+ g(x)$, will be used to denote that there is a constant c such that for all x allowed in the context of the expression, one has $f(x) \leqslant_+ g(x)$. Such a constant is allowed to depend on the initial choice of oupi-interpreter ϕ, but not on any other parameter.

More formally this notation is defined as follows. Let $QuRx$ denote a series $Q_1u_1Q_2u_2 \ldots Q_ku_kRx$ of quantifiers $Q_i, R \in \{\forall, \exists\}$ over the variables $u_i, i \leqslant k$,

that are implicitly or explicitly stated in the context of an equation. The expression

$$QuRx\Big[f_u(x) \leqslant_+ g_u(x)\Big]$$

means

$$\exists c QuRx\Big[f_u(x) \leqslant g_u(x) + c\Big].$$

Since Kolmogorov complexity is dependent on the choice of ϕ, and since often a fixed amount of instructions are added to some program generating x, bounds on Kolmogorov complexity $K(x)$ are typically expressed within $\leqslant_+$ bounds.

By universality of ϕ, two main strategies exist to show that some $x \in 2^{<\omega}$, satisfies $K(x) \leqslant_+ k$.

1. Write a program for some prefix-free Turing machine (or your favorite programming language) producing x of length shorter than k.
2. Define a partial computable function that is defined on a prefix-free set of $2^{<\omega}$, and outputs x on input p with $l(p) \leqslant k$.

Some well known bounds for Kolmogorov complexity are now shown for suitable ϕ using the second strategy. Let $n \in \omega$, $x \in 2^n$ and $y \in 2^{<\omega}$.

1. $$K_0(n) \leqslant_+ n$$

   Proof. Let

   $$F_0(\overbrace{0\ldots0}^{n}1) = n.$$

   Remark that F_0 is computable and defined on a prefix-free set.

2. $$K_0(x|n) \leqslant_+ n$$

   Proof. Let

   $$G_n(y) = \begin{cases} y & \text{if } y \in 2^n \\ \infty & \text{otherwise.} \end{cases}$$

   Remark that for each n, the function G_n is prefix-free.

3. For suitable ϕ

   $$K_s(x, y) \leqslant_+ K_s(x) + K_s(y|x).$$

   Proof. A well-defined command w can be added on ϕ such that

   $$\phi_s(wpq) \downarrow = [\phi_s(p), \phi_s(p|\phi(q))],$$

   when defined, and undefined otherwise.

4. For suitable ϕ
$$K(x) \leqslant_+ n + 2\log n.$$
Proof. Remark that the decimal expansion of n belongs to $2^{\log(n+1)}$.
$$K(x) \leqslant_+ K(x|n)+K(n|\log n)+K(\log n|\log\log n) \leqslant_+ n+\log n+\log n.$$

5. For suitable ϕ
$$K_s(x) =_+ K_s(K_s(x), x).$$
Proof. Let a well-defined command w be added on ϕ such that
$$\phi_s(wp|y) = [\phi_s(p|y), p],$$
when defined, and undefined otherwise.

A few more observations for $x, y \in \omega^{<\omega}$ or $x, y \in 2^{<\omega}$:
$$\begin{aligned} K(x|x) &\leqslant_+ 0 \\ K(xx) &=_+ K(x) \\ K(xy) &\leqslant_+ K(x) + K(y). \end{aligned}$$
For $x, y \in \omega$
$$K(x + y) \leqslant_+ K(x) + K(y).$$

Theorem 2.4.2. *For every n there is at least one $x \in 2^n$ such that $K(x|n) \geqslant n$.*

Proof. The amount of descriptions on a binary tape of length less than n are upper bounded by:
$$2^0 + 2^1 + 2^2 + ... + 2^{n-1} = 2^n - 1.$$
On the other hand, there are 2^n different strings in 2^n. Therefore, at least one string in 2^n must not have a shortest description larger than n. □

2.4.2 Coding theorem

Theorem 2.4.3 (Shannon-Fano Code)**.**

(i) Let P be a Δ_1-semimeasure then there is a prefix-free code E with $-\log P(x) < l(E(x)) + 1$.

(ii) Let P be a Σ-semimeasure then there is a prefix-free code E with $-\log P(x) \leqslant l(E(x)) + 3$.[3]

[3] Remark that this constant 3 can be improved to 2 [44].

(iii) For suitable ϕ, if a program p satisfies $\phi_s(p|x) \downarrow = -\log P(x)$ for all $x \leqslant s$, then for such x: $K_{s+1}(x) \leqslant_+ -\log P(x) + l(p)$.

Proof.
(i) Let $Pc(x) = \sum\{P(y) : l(y) = n \wedge y < x\}$. Then $E(x)$ is given by the decimal expansion of $Pc(x)$ truncated at the length $l(E(x))$ that satisfies $-\log P(x) \leqslant l(E(x)) < -\log P(x) + 1$. Remark that the range of $E(x)$ is prefix-free. To compute x from $E(x)$, one computes $Pc(y)$ for increasing y until the first $Pc(y)$ is found for which the decimal expansion corresponds to $E(x)$.
(ii) Let

$$Z = \{(x,k) : P_0(x)2^k \leqslant P(x)\}.$$

The set Z is enumerable, and let (x_i, k_i), for $i \in \omega$ be the corresponding enumeration. Define $Q(i) = P_0(x_i)2^{k_i - 1}$. Remark that $\max_i\{Q(i) : x_i = x\} \geqslant P(x)/4$, and remark that $Q(i)$ defines a semimeasure:

$$\sum_i Q(i) \leqslant \sum_x P_0(x) \sum\{2^{k-1} : (x,k) \in Z\} \leqslant \sum P(x) \leqslant 1$$

Apply the coding strategy of (i) on $Q(i)$, and let $E(x)$ be the code of $\arg\max\{k_i : x_i = x\}$.
(iii) Remark that the decoding of $E(x)$ requires an evaluation of all $-\log P(y)$ for increasing $y \leqslant s$, and the corresponding evaluation of all codes $E(y)$, until a code $E(y) = E(x)$ is found. This procedure shows that the following assignment defines a well-defined command:

$$\phi_s(wpE(x)) \downarrow = x,$$

for all $x \leqslant s$, and $\phi_s(wpE(x)) = \infty$ otherwise. □

The following result is known as the Coding Theorem [44].

Theorem 2.4.4. *For any universal Σ-semimeasure m with $K(m) \leqslant_+ 0$:*

$$-\log m(x) =_+ K(x) + c.$$

Proof. The Kraft inequality [44] states that for any prefix-free set S of binary strings:

$$\sum\{2^{-l(p)} : p \in S\} \leqslant 1.$$

Therefore $Q_K(x) = 2^{-K(x)}$ defines a semimeasure. Remark that $K(x)$ is a Π-function, and therefore Q_K is a Σ-semimeasure. By universality of m it follows that $-\log m(x) \geqslant_+ K(x)$, and because m is a Σ-semimeasure, the Coding Theorem 2.4.3 can be applied, showing that

$$-\log m(x) \geqslant_+ K(x).$$

□

Two universal Σ-semimeasures will be often used:

$$Q_p(x) = \sum\{2^{-l(p)} : \phi(p) \downarrow = x\} \tag{2.5}$$

$$Q_K(x) = 2^{-K(x)}. \tag{2.6}$$

A bivariate universal Σ-semimeasure can also be defined, and it follows in the same way that $-\log m(x,y) =_+ K(x,y)$.

2.4.3 Additivity of Kolmogorov complexity

Proposition 2.4.5. *For suitable ϕ and $s \geqslant l(y)$, one has*

$$K_{s-1}(x,y) \geqslant_+ K_s(x) + K_s(y|x, K_s(x)).$$

The proof is essentially the same as additivity of prefix-free Kolmogorov complexity [44], but formulated with indexes.

Proof. Let for some constant c large enough

$$S_{s,x} = \{p : \phi_s(p) \downarrow = [x,z] \wedge l(p), l(z) \leqslant 2s + 2l(x)\}.$$

Remark that for any x: $S_{\infty,x}$ can be enumerated from x. Therefore,

$$Q_s(x) = \sum\{2^{-l(p)} : p \in S_{s-1,x}\}$$

is a lower semicomputable semimeasure. By the Coding Theorem:

$$K_s(x) \leqslant_+ -\log \sum\{2^{-l(p)} : p \in S_{s-1,x}\}.$$

Let,

$$P_{s-1}(z) = 2^{K_s(x)-O(1)} \sum\{2^{-l(p)} : \phi_{s-1}(p) \downarrow = [x,z]\},$$

define a conditional semimeasure that can be computed from $x, K_s(x)$. The evaluation of $P_{s-1}(y)$ for all y with $l(y) \leqslant s$ can be combined with Shannon Fano code, in a well-defined command to obtain for suitable ϕ that

$$K_s(y|x, K_s(x)) \leqslant_+ K_{s-1}(x,y) - K_s(x).$$

□

Corollary 2.4.6.

$$K(x,y) =_+ K(x) + K(y|x, K(x))$$

2.4.4 Equivalence of shortest programs for an x and the "computes" relation

For any $a, b \in \omega^{<\omega}$, the expression "$a \longrightarrow b$", means that for any a, b allowed in the context of the expression, b can be computed from a by some fixed instructions. This is formalized in an analogues way as the definition of the $\leqslant_+$ operator in Subsection 2.4.1. Let $QuRx$ denote a series $Q_1u_1Q_2u_2 \ldots Q_ku_kRx$ of quantifiers $Q_i, R \in \{\forall, \exists\}$ over the variables u_i, $i \leqslant k$, that are implicitly or explicitly stated in the context of an equation. The expression

$$QuRx\Big[a_u(x) \longrightarrow b_u(x)\Big]$$

means that

$$\exists f \in \Delta_1 QuRx\Big[f(a_u) = b_u\Big].$$

Remark that for any context $a \longrightarrow b$ implies $K(b) \leqslant_+ K(a)$, and $K(b|a) \leqslant_+ 0$. It is said that a *computes* b (notation "$a \longrightarrow_+ b$") iff $K(b|a) \leqslant_+ 0$. Remark that the program that computes b from a, might differ for different a's allowed within the context of the statement, however the length of the program must be bounded by a constant that is indifferent from the choice of a. To make the reader used to this difference, and for later reference, some more results are given.

Theorem 2.4.7 (Gacs[4]). *For $x^* = \min\{p : \phi(p) \downarrow = x\}$, and p such that $\phi(p) \downarrow = x$ and $l(p) \leqslant_+ K(x)$, one has*

$$x, K(x) \longleftrightarrow_+ p \longleftrightarrow_+ x^*$$

Proof. Obviously, $x^* \longrightarrow_+ K(x)$ and $p \longrightarrow_+ K(x)$. It suffices to show that $x, K(x) \longrightarrow_+ p$. All programs p satisfying the conditions of the Theorem can be enumerated from x and $K(x)$. By the Coding Theorem applied to $m = Q_p$ (Equation (2.5)) one observes that the amount c of possible programs p satisfying the conditions of the theorem, satisfies $c \leqslant_+ 0$. Therefore, p is computed from $x, K(x)$ by a fixed amount of instructions generating this enumeration, and by the corresponding index of p. □

Corollary 2.4.8.

$$K(x, y), x, y \longrightarrow_+ K(x), K(x|y^*)$$

Proof. By additivity of Kolmogorov complexity. □

[4] In the Second Edition of [44, Ex. 3.7.7] a more general result is stated with a more standard notation. The result is attributed to [25]

For the record we remark that the following proposition can be shown.

Proposition 2.4.9. *For some ϕ one has*

$$K(x), x \longrightarrow x^*$$

For some ϕ there are infinitely many x such that

$$K(x), x \not\longrightarrow x^*$$

However, for Corollary 2.4.8 the improved result is an open question.

Question 2.4.10. *Is there an oupi-interpreter such that*

$$K(x, y), x, y \longrightarrow K(x), K(x|y^*)?$$

3

Minimal typical models and sophistication

Abstract. m-sophistication and explicit minimal typical models are defined for use in technical proofs the subsequent chapters. The relation with the literature is shown. m-sophistication:

- defines a sumtest for a universal semimeasure,
- is the main tool in the proof of [23] to show that high complexity of complexity is rare,
- is within small additive terms lower than sophistication, and larger than coarse sophistication,
- equals within small terms the length of a probabilistic minimal typical model.

The principle of Occam's razor, applied to model selection, can be interpreted in two different ways, which lead to the definitions of minimal sufficient statistics, and minimal typical models. It is shown that for strings of small length, under some stability conditions, a minimal sufficient statistic is equivalent to a minimal typical model, and vice versa. Furthermore these models are equivalent to initial segments of the Halting probability, and thus contain mostly Halting information. However, in a strict mathematical way, for large strings, both formalizations may lead to different solutions, since it is shown that most information contained in a minimal typical model is equivalent with an initial segment of the Halting probability, while minimal sufficient statistics can carry a substantial amount of non-Halting information. Weak sufficient statistics are introduced, and it is shown that minimal weak sufficient statistics are equivalent to minimal typical models.

The definition of "typical models" is originally motivated by the first hypothesis testing question, and is given in Subsection 3.3.3. However for relation with hypotheses testing, sumtests are more suitable. The chapter owes its motivation to the introduction of m-sophistication for later use, and to important questions on model selection.

3.1 m-sophistication

The Kolmogorov complexity of a finite binary sequence is a measure for the amount of structure in a finite discrete sequence. Sophistication [3, 39] is a measure to quantify the complexity of this structure. m-sophistication is a variation of sophistication. The concept of m-sophistication was used in the proof that high complexity of complexity is rare [23, 25, 44]. m-sophistication is defined relative to a universal semimeasure m. In future chapters it will be often applied to the universal semimeasures

$$\begin{aligned} Q_p(x) &= \sum\{2^{-l(p)} : \phi(p) \downarrow = x\} \\ Q_K(x) &= 2^{-K(x)}. \end{aligned}$$

3.1.1 Halting probability and a Buzzy Beaver variant

In computability theory, the number Ω is typically defined as the prior probability that some universal prefix-free Turing machine halts if the input string in 2^∞ is drawn from a uniform distribution [11, 23]. Here a closely related concept is studied: the a prior probability that a universal semimeasure is defined. Lemma's and propositions are derived for later use, and connected to a dominance relation for Real numbers [65].

Definition 3.1.1. Let m be some universal semimeasure.

$$\begin{aligned} \Omega_{m,t} &= \sum_{x<t} m_t(x) \\ \Omega_m &= \lim_{t\to\infty} \Omega_t \end{aligned}$$

The original definition in [11, 23] is obtained by choosing $m = Q_p$, as in equation 2.5. For $\alpha \in 2^\omega$ the notation $\alpha^n = \alpha_1 \ldots \alpha_n$ is used. Ω_{Q_p} satisfies the following well known theorem.

Theorem 3.1.2. *For all n: $K(\Omega^n_{Q_p}) \geqslant_+ n$. There is a constant c such that for all n, the Halting of any program $p \in 2^{<n-c}$ can be decided by $\Omega^n_{Q_p}$.*

It will be shown later in this section that these properties of Ω_{Q_p} remain for general Ω_m with a similar argument.

For some $\alpha, \beta \in 2^\omega$, and for all n, the equivalence relation $\alpha^n \longrightarrow_+ \beta^n$ defines a partial order on 2^∞. This order is equivalent with the 'domination' relation introduced in [65], and also used in [9][1]. Ω_{Q_p} remains in the same equivalence class for several choices of universal machine ϕ. Let ϕ and ϕ' be two optimal universal prefix-free Turing machines, and let Q_p and Q'_p be defined as in equation 2.5 relative to ϕ and ϕ', then it is easily observed that

$$\Omega^n_{Q_p} \longrightarrow_+ \Omega^n_{Q'_p}.$$

Another example of such a relation is

$$\Omega^n_{Q_p} \longrightarrow_+ \Omega^n_{Q_K},$$

where Q_K is defined in equation 2.6. It is an interesting question whether the opposite direction also holds.

Question 3.1.3.

$$\Omega^n_{Q_p} \longrightarrow_+ \Omega^n_{Q_K}$$

Following the proof that high $K(K(x)|x)$ is rare in [25], the times t_n are defined. *Fix* some universal semimeasure m for this subsection, and assume $K(m) \leqslant_+ 0$.

Definition 3.1.4. For each n let

$$t_n = \min\{t : \Omega_m - \Omega_{m,t} \leqslant 2^{-n}\}.$$

It is easily observed that

Lemma 3.1.5.

$$\Omega^n_m \longleftrightarrow_+ n, t_n.$$

Remind the definition of $t[p]$ in Subsection 2.3.2.

Lemma 3.1.6. *For c large enough, and any halting $p \in 2^{<\omega}$:*

$$\begin{aligned} \phi(p) &\leqslant t_{l(p)+c} \\ t[p] &\leqslant t_{l(p)+c}. \end{aligned}$$

Proof. Let $n = l(p) + c + 1$ with c large enough, and Let $x \in 2^{<\omega}$ be the lexicographic first string with $-\log m_{t_n}(x) \geqslant l(x) \geqslant 2n$. Suppose that $\phi(p) \geqslant t_n$, then $p \longrightarrow_+ p, n \longrightarrow_+ x$ and thus

$$-\log m(x) \leqslant_+ K(x) \leqslant_+ l(p).$$

[1] Strictly mathematically, one should define this domination relation using the " $\longrightarrow^+$ " notation as described in Section 4.1, to obtain an equivalence in all possible contexts.

which implies for c sufficiently large

$$\Omega_m - \Omega_{m,t_n} \geqslant m(x) - m_{t_n}(x) \geqslant 2^{-l(p)-c} - 2^{-2n} > 2^{-l(p)-c-1},$$

contradicting the definition of t_n. The second claim follows by remarking that for every Halting p: $p \longrightarrow t[p]$. □

The prefix-free Buzzy Beaver function is defined by:

$$PBB(n) = \max\{\phi(p) : l(p) \leqslant n\}.$$

Lemma 3.1.7 shows that t_n is a very fast growing function that oscillates between $PBB(n)$ and $PBB(n + 2\log n)$

Lemma 3.1.7. *For all n*

$$n \leqslant_+ K(t_n) \leqslant_+ n + 2\log n.$$

There exists a constant c such that:

$$PBB(n - c) \leqslant t_n < PBB(n + 2\log n + c).$$

Proof. From Lemma 3.1.6, it follows that

$$PBB(n - c) \leqslant t_n$$

and thus also

$$K(t_n) \geqslant_+ n$$

By Lemma 3.1.5 one has

$$K(t_n) \leqslant_+ K(\Omega_m^n) \leqslant_+ n + K(n).$$

□

The dependence of t_n on the choice of m is given by the subsequent corollary.

Corollary 3.1.8. *For all universal semimeasures m, and m', there is some constant c such that*

$$t_n < t'_{n+2\log n+c},$$

with t_n and t'_n defined by m and m'.

Proof.

$$t_n \leqslant PBB(n + 2\log n + c) < t'_{n+2\log n+2c}$$

□

A Real number $\alpha \in 2^\omega$ is *(Martin-Löf) random* if there is a constant c, such that for any n: $K(\alpha^n) \geqslant n - c$. For more background on randomness for Real numbers it is referred to [18, 48]. It follows by Lemma 3.1.7 that

Corollary 3.1.9. *Ω_m is random.*

Proof. Since $n \leqslant_+ K(t_n) \leqslant_+ K(\Omega_m^n)$. □

Remark that by this result, and by [9] it follows that the set of all Ω_m with m universal corresponds to all computably Σ-random Real numbers. From Corollary 3.1.8 the subsequent lemma is proved.

Lemma 3.1.10. *for m, and m' universal semimeasures there exists a c such that*

$$\Omega_m^n \longrightarrow \Omega_{m'}^{n-2\log n-c}.$$

Proof. For m, and m' such that $K(m) \leqslant_+ 0$, and $K(m') \leqslant_+ 0$, one has

$$\Omega_m^n \longrightarrow_+ n, t_n \longrightarrow_+ n, t'_{n-2\log n} \longrightarrow_+ \Omega_{m'}^{n-2\log n}.$$

□

The question rises, whether the set of all Ω_m for some universal semimeasures has a maximal element relative to the partial $\longrightarrow_+$ order. This is equivalent with to the question whether there is a maximal computably Σ-random Real for the partial $\longrightarrow_+$ order.

Finally, it can be asked how tight the bounds of Lemma 3.1.7 are. Firstly, remark that for a random $\alpha \in 2^\omega$ only a small amount of values $K(\alpha^n)$ are allowed:

$$n \leqslant_+ K(\alpha^n) \leqslant_+ n + 2\log n.$$

It is well known that $K(\alpha^n)$ oscillates within these bounds.

Lemma 3.1.11. *For any random $\alpha \in 2^\omega$ there is a constant c such that there are an infinite amount of n with*

$$K(\alpha^n) \leqslant n + 2\log\log n + c,$$

and there are an infinite amount of n such that

$$K(\alpha^n) \geqslant n + \log n - c.$$

Proof. For any k, let n be such that the decimal expansion of n equals $1\alpha^k$. Remark that $\alpha^k \longrightarrow n$, and $\log n =_+ k$. Remark that for any $z \in 2^{n-k}$ one has

$$K(z|\alpha^k) - c \leqslant K(z|n-k) \leqslant_+ n - k + c,$$

and consequently,

$$\begin{aligned} K(\alpha^n) &\leqslant_+ K(\alpha^k) + K(\alpha_{k\ldots n}|\alpha^k) + c \\ &\leqslant_+ k + 2\log k + n - k. \end{aligned}$$

The second inequality follows from Exercise 3.6.3d in [44]. □

This observation can be used to show a similar result for Lemma 3.1.7, however the proof is not so difficult, but rather long, and is therefore omitted.

3.1.2 m-sophistication and complexity of complexity

Definition 3.1.12. For any universal semimeasure m, for $c \in \omega$, and for some $x \in 2^{<\omega}$, the *m-sophistication* of x is

$$k_c(x) = \min\{k : K_{t_k}(x) \leqslant K(x) + c\}.$$

$k_c(x)$ is limit-computable in x, but not lower semicomputable or upper semicomputable, even within large error, by Proposition 3.1.20. From Corollary 3.1.8 it is observed that k_c is relatively stable with respect to changes of universal semimeasure m.

Corollary 3.1.13. *Let m and m' be universal semimeasures and let k and k' be the m-sophistication and m'-sophistication, then there is a c' such that for any c:*

$$k_c \leqslant k'_c + 2\log k'_c + c'.$$

As for sophistication (see further), also m-sophistication is unstable with respect to the parameter c. Remind that in this chapter m is assumed to be fixed such that $K(m) \leqslant_+ 0$.

Lemma 3.1.14. *For all c, there is a c' such that for infinitely many x:*

$$k_c(x) - k_{c+c'}(x) \geqslant_+ l(x) - 4\log l(x).$$

Informally, one chooses an x that is only a little compressible, by some constant $c + c'$, for c' large enough. Thus, $k_{c+c'} = 0$. However, this little compression only appears after a time $t_{n-O(\log n)}$. Therefore, k_c is much larger. A formal detailed proof needs some care, and is given below.

Proof. Let n be short for $l(x)$. The Proposition follows by showing that that for all c, there is a c' such that for infinitely many x: and $k_{c+c'}(x) \leqslant_+ 2\log n$, and $k_c(x) \geqslant_+ n - 2\log n$. This follows for some computable function f by

$$\begin{aligned} K(x) &\geqslant n - c - c' \\ K_{f(n)}(x) &\leqslant_+ n \\ K(x) &\leqslant_+ n - c \\ K_{t_{n-2\log n-c-c'}}(x) &\geqslant n - 1. \end{aligned}$$

By the existence of Solovay functions [65], it can be shown that there is a function f such that for infinitely many n one has $K_{f(n)}(n) = K(n)$. This shows that

$$\begin{aligned} K(x) &=_+ K(x,n) \\ &=_+ K(x|n^*) + K(n) \\ &=_+ K(x|n) + K(n), \end{aligned}$$

and also corresponding time bounded versions hold. From this it follows that it suffices to show that for each n there exists an $x \in 2^n$ such that

$$\begin{aligned} K(x|n) &\geqslant n - c - c' \\ K_{f(n)}(x|n) &\leqslant_+ n \\ K(x|n) &\leqslant_+ n - c \\ K_{t_{n-2\log n-c-c'}}(x|n) &\geqslant n - 1. \end{aligned}$$

Remark that the second equation of (3.1.2) follows directly. Kolmogorov complexity fluctuates "continuously", in the sense that there exists a constant e such that for all a, r abs$(K(r+1,a) - K(r,a)) \leqslant e$. Since for all n: $K(t_{n-K(n)-c-2e}) \leqslant n - c - e$, there always exists an r such that:

$$n - c - 2e \leqslant K(r, n, t_{n-K(n)-c-2e}) < n - c - e.$$

Remark that $r \leqslant n^2 2^e$ can be chosen for n large enough. Let $t = t_{n-K(n)-c-2e}$, and let $x \in 2^n$ the lexicographic r-th string in 2^n such that $K_t(x|n) \geqslant n - 1$. Remark that the forth equation of (3.1.2) is satisfied. Remark that there are enough $x \in 2^n$ that satisfy this condition. Also, remark that

$$t, r, n \longleftrightarrow_+ x.$$

This implies that for e large enough:

$$n - c - 3e \leqslant K(x|n) < n - c.$$

Therefore

$$c < K_t(x) - K(x) \leqslant c + 3e.$$

This shows the first and third equations in (3.1.2). □

Inspired by the Coding Theorem, a definition very related to m-sophistication is given by (m,m)-sophistication:

$$k'(x) = \min\{k : \frac{m(x)}{m_{t_k}(x)} \leqslant 2\}.$$

Lemma 3.1.15. *For any c large enough: $k' \geqslant_+ k_c$.*

Proof. For suitable ϕ the Coding Theorem implies

$$K_{t_{k'(x)}+1}(x) \leqslant_+ -\log m_{t_{k'(x)}}(x) =_+ -\log m(x) =_+ K(x).$$

For general ϕ the Coding Theorem implies that there is a constant c large enough such that

$$K_{t_{k'(x)+c}}(x) \leqslant_+ K(x).$$

□

High (m, m)-sophistication is rare.

Lemma 3.1.16. *For any k and $S_k = \{x : k'(x) \geqslant k\}$:*

$$m(S_k) \leqslant 2^{-k+1}.$$

Proof.

$$\frac{1}{2}m(S_k) \leqslant m(S_k) - m_{t_k}(S_k) \leqslant \Omega - \Omega_{t_k} \leqslant 2^{-k}.$$

□

Lemma 3.1.17. *Let $k(x)$ be either $k'(x)$ or $k_c(x)$ for any c, then:*

$$K(K(x)|x) \leqslant_+ k(x) + 2\log k(x).$$

Proof. Remark that $t_{k(x)}, x \longrightarrow_+ K(x)$, thus

$$K(K(x)|x) \leqslant_+ K(t_{k(x)}) \leqslant_+ K(\Omega^{k(x)}) \leqslant_+ k(x) + 2\log k(x),$$

where the last inequality follows from Lemmas 3.1.5 and 3.1.11. □

Now it easily follows from Lemmas 3.1.16 and 3.1.17 that high complexity of complexity is rare [23].

Corollary 3.1.18. *There exists a constant $c > 0$ such that*

$$m(\{K(K(x)|x) \geqslant k\}) \leqslant c2^{-k-2\log k}.$$

It also follows from Lemma 3.1.16 that (m, m)-sophistication and m-sophistication define sumtests for m. Remark that this gives an extra interpretation of sophistication in terms of randomness deficiency, additional to those in [4].

Corollary 3.1.19. *For $k = k'$ and for $k = k_c$ with c large enough, $k - 2\log k$ defines a sumtest for m.*

Proof.

$$\sum_{x\in 2^{<\omega}} m(x)2^{k(x)-2\log k(x)-2} \leqslant \sum_{k\in\omega} m(S_k)2^{k-2\log k-2} \leqslant \sum_{k\in\omega} 2^{-2\log k-1} \leqslant 1$$

□

k_c and k' are not computable, and not even a logarithmic lower bound can be computed by the following proposition.

Proposition 3.1.20. *For $k = k'$ and for $k = k_c$ with c large enough, k can not be approximated by a lower or upper semicomputable function within $k - 2\log k + O(1)$ error.*

Proof. Suppose that the function d approximates k such that $k - d \leqslant k - e\log k + O(1)$ for some constant e. This implies that $d \geqslant e\log k - O(1)$. Remark that this implies by Corollary 3.1.19 that there exists a c' such that $d - 4\log d - c'$ is a sumtest for m.

By Proposition 4.1.5 every lower semicomputable sumtest for m is bounded by a constant, which implies that if d was lower semicomputable, then $d \leqslant_+ 0$, and thus only the constant $e = 0$ is allowed.

By Corollary 4.2.5 every upper semicomputable sumtest for m is bounded by $\log l(x) + O(\log\log l(x))$. Therefore, only the constant $e = 1$ is allowed. □

3.1.3 Sufficient statistics, sophistication and coarse sophistication

A binary string $x \in 2^n$ can be said to have structure or regularities if $K(x) < n - c$. Many notions of sophistication and computational depth, express how "sophisticated" this structure is in several contexts [1, 2, 4, 35].

Definition 3.1.21. Let f be a computable function. A function *f-sufficient statistic* [26] is a computable prefix-free function g such that there exists a $d \in g^{-1}(x)$ with

$$K(g) + l(d) \leqslant K(x) + f(l(x)).$$

For some constant c, a c-sufficient statistic is an f-sufficient statistic for which f is the constant c function.

Definition 3.1.22. The sophistication [39] of $x \in 2^{<\omega}$ is given by:

$$k_c^{soph}(x) = \min\{K(f) : f \text{ is a } c\text{-sufficient statistic of } x\}.$$

Remark that there is a slight deviation from [1, 3, 39, 71], since it is also required that f is prefix-free. This is necessary to interpret sophistication as the length of a minimal sufficient statistic [26]. Also, remark that now Lemma 3.1.25 is true. Let $bb(x)$ be the inverse of the prefix-free Buzzy Beaver function, it is $bb(x) = \min\{k : x \leqslant PBB(k)\}$. It is a very slow growing function, dominated by any unbounded non-decreasing function [7]. The following proposition

is closely related to the equality of Buzzy Beaver depth and coarse sophistication within logarithmic terms in $l(x)$ [3].

Proposition 3.1.23. *There exists a c' such that for all c, x:*

$$k_{c+c'}(x) \leqslant_+ k_c^{soph}(x) + bb(x).$$

Proof. The inequality follows by observing that any function f, witnessing the definition of the m-sophistication $k_c^{soph}(x)$, defines a prefix-free description of x of length $K(x) + c + c'$, for some c' large enough. Let $d = \min\{d : f(x) = d\}$, let

$$M = PBB(bb(x)) \geqslant x \geqslant d,$$

and let p be the program that evaluates all these possible programs using $f(e)$ for all $e \leqslant M$. Let $s = t[p]$ be the computation "time" of these evaluations. Remark that $K_s(x) \leqslant K(x) + c + c'$, and thus

$$s \geqslant t_{k_{c+c'}(x)} \geqslant PBB(k_{c+c'} - c')$$

for some c' large enough, by Lemma 3.1.7. This implies

$$k_{c+c'} \leqslant_+ K(s) \leqslant_+ l(p) \leqslant_+ K(f) + bb(x) \leqslant_+ k^{soph}(x) + bb(x).$$

□

Definition 3.1.24. A *probabilistic f-sufficient statistic* is a measure P such that

$$K(P) - \log P(x) \leqslant K(x) + f(l(x)).$$

Since prefix-free functions are used here, probabilistic and function sufficient statistics are equivalent.

Proposition 3.1.25. *There is a constant c such that every probabilistic f-sufficient statistic P defines a function $(f + c)$-sufficient statistic g with $P \longleftrightarrow g$. There is a constant c such that for any f with $f(x) \leqslant x$ every function f-sufficient statistic g defines a probabilistic $(f + c)$-sufficient statistic P with $P \longleftrightarrow g$.*

Proof. The first claim follows by the same proof as in [71]. It remains to show the second claim. Let g be the function f-sufficient statistic, and let

$$P(x) = \sum\{2^{-l(d)} : g(d) = x \wedge d \leqslant x2^x\}.$$

Remark that $P(x) = 0$ if there is no $d \leqslant x$ with $g(d) = x$. It follows that $-\log P(x) \leqslant l(d)$ for the witness d of x in the definition of the function f-sufficient statistic of g. Remark that $K(g) \leqslant_+ K(P)$, and therefore the conditions of the definition of $(f + c)$-sufficient statistic are fulfilled. □

Let

$$P_k(x) = N2^{-k}(m_{t_k}(x) - m_{t_{k-1}}(x)), \tag{3.1}$$

where N is a normalization constant such that P_k defines a computable measure. Remark that $2 \leqslant N < 4$. Also, remark that this can be considered as the probabilistic equivalent of the "explicit minimal near sufficient set statistic" described in [26].

Lemma 3.1.26. *For $m = Q_K$:*

$$K(x|\Omega^{k'(x)}) \leqslant_+ K(x) - k'(x).$$

Proof. Remark that since $m = Q_K$, for any k either $m_{t_k}(x) = m_{t_{k-1}}(x)$ or $m_{t_k}(x) = 2m_{t_{k-1}}(x)$. This implies that $P_{k'(x)}(x) = 2^{-K(x)-1+k'(x)}$. The Lemma follows by Shannon-Fano coding. □

Proposition 3.1.27. *For $m = Q_K$*

$$k'(x) \leqslant_+ K(x) - K(x|\Omega^{k'(x)}) \leqslant_+ k'(x) + 2\log k'(x).$$

Proof. Lemma 3.1.26 shows the left inequality, and the right inequality follows by additivity of Kolmogorov complexity, and $K(\Omega^n) \leqslant_+ n + 2\log n$. □

To relate P_k to sophistication, it is shown that it defines some f-sufficient statistic.

Proposition 3.1.28. *There exists a c such that*

(i) for $m = Q_K$, $P_{k'(x)}$ is a probabilistic $(2\log k'(x) + c)$-sufficient statistic for x,

(ii) for $m = Q_K$, and for any c, $P_{k_c(x)}$ is a $(2\log k_c(x) + c + c')$-sufficient statistic for x.

(iii) for any m and any c', there is a $k \leqslant_+ k_{c'}(x)$ such that P_k is a $(3\log k_c(x) + c + c')$-sufficient statistic for x.

Proof.
(i). Remark that for any k: $K(P_k) \leqslant_+ k + 2\log k$. Choosing $k = k'(x)$, and by Lemma 3.1.26, the result follows.
(ii).

$$m_{t_{k_c(x)}}(x) - m_{t_{k_c(x)-1}}(x) \geqslant \frac{1}{2} m_{t_{k_c(x)}}(x) \geqslant 2^{-c-1} m(x).$$

This shows that

$$-\log P(x) \leqslant_+ K(x) - k_c(x).$$

(iii). By some time bounded version of the Coding Theorem there is a constant e such that:

$$\log m(x) =_+ K(x) \leqslant K_{t_{k_c}}(x) + c \leqslant_+ -\log m_{t_{k_c(x)+e}}(x).$$

Therefore

$$m(x) \leqslant^* m_{t_{k_c(x)+e}}(x) =^* \sum\{2^k P_k(x) : k \leqslant k_c(x) + e\}.$$

This shows that there is for every x an $k \leqslant k_c(x) + e$ such that

$$m(x) \leqslant^* \frac{2^k}{k} P_k(x).$$

By applying the Coding Theorem, and taking $-\log$ of the above equation one obtains:

$$\begin{aligned} K(x) &=_+ -\log m(x) \\ &\geqslant_+ k - \log k - \log P_k(x) \\ &\geqslant_+ K(P) - 3\log k - \log P_k(x). \end{aligned}$$

Which shows that P_k is a $(3\log k + e')$-sufficient statistic. Remark that $e' \leqslant c + c'$ for some c' independent of c. □

By a similar argument as in 3.1.14 for m-sophistication, sophistication is unstable with respect to the parameter c, therefore coarse sophistication [3] is defined as

$$k^{csoph}(x) = \min_c\{k_c(x) + c\}.$$

As a corollary of Proposition 3.1.28 it follows that:

Corollary 3.1.29. *For any c*

$$k^{csoph}(x) \leqslant_+ k_c(x) + 2\log k_c(x).$$

Proposition 3.1.30. *$k^{csoph}(x) - 4\log k^{csoph}(x)$ defines a sumtest for m. k^{csoph} can not be approximated by a lower or upper semicomputable function within $k - 2\log k + O(1)$ error.*

Proof. This follows from Corollary 3.1.29 and the same proof as Proposition 3.1.20. □

Finally, it is remarked that bivariate sophistication $k_c(x, y)$ and conditional sophistication $k(x|y)$ can be defined. It can be shown[2] that

$$\forall c \exists c' \Big[k_{c+c'}(x, y) - O(\log k_{c+c'}) \leqslant k_c(x) + k_c(y|x, K(x)) \leqslant k_{c-c'}(x) + k_{c-c'}(y|x, K(x))\Big].$$

[2]The proof is not so difficult but rather long, and will be written out when the author has time for it. Hint: use $m = Q_p$, and derive some additivity result for the PBB function. Let me know if you have written this out.

3.2 Probabilistic minimal sufficient statistics and initial segments of Ω

The principle of Occham's Razor[3] can be interpreted as preferring the smallest set of rules to reach a specific goal. For model selection, one looks for the smallest model that in some way captures all regularities of the data. A function model M defines for each data x, some r such that $x = M(r)$. There are two plausible ways to state that a model M captures all regularities in x:

- One can not give a shorter description for r given M then some natural encryption of r (Typical model).
- One can not give a shorter description for x then a shortest description for M and a natural encryption of r (Sufficient statics).

By the principle of Occam's razor, one is interested in the minimal typical models, and the minimal sufficient statistics. Remark that model selection through the minimal sufficient statistic is also called "Minimum Description Length principle" [28]. Two central questions are now investigated in the remaining of this chapter:

- Is a minimal typical model for some x equivalent with a minimal sufficient statistic of x ?
- Is a minimal typical model, or a minimal sufficient statistic equivalent with some initial segment of a Halting probability Ω_m ?

From now on, "sufficient statistic" is abbreviated as SS, and "minimal SS" as MSS.

3.2.1 $P_{k_c(x)}$ can be almost computed from any probabilistic c-MSS

Proposition 3.1.27 shows that any x contains some amount of information of the Halting probability, which allows by Proposition 3.1.28 to contain all necessary information to compress x. Proposition 3.1.28 further elaborates this observation, and defines some near probabilistic sufficient statistic. It is often stated that for many practical applications, one can assume that a logarithm is upper bounded by a constant. This conjecture also applies here, since for any practical datasets x,

[3] *The term razor refers to the act of shaving away unnecessary assumptions to get to the simplest explanation. No doubt this maxim represents correctly the general tendency of his[William Occham] philosophy, but it has not so far been found in any of his writings. His nearest pronouncement seems to be Numquam ponenda est pluralitas sine necessitate [Plurality must never be posited without necessity], which occurs in his theological work on the Sentences of Peter Lombard (Quaestiones et decisiones in quattuor libros Sententiarum Petri Lombardi (ed. Lugd., 1495), i, dist. 27, qu. 2, K). In his Summa Totius Logicae, i. 12, Ockham cites the principle of economy, Frustra fit per plura quod potest fieri per pauciora [It is futile to do with more things that which can be done with fewer].*
Thorburn, 1918, pp. 352-3; Kneale and Kneale, 1962, p. 243.

one has that $l(x) \leqslant 2^{50}$, therefore, if arbitrary constants of length 50 are allowed, one has $\log k'(x) \leqslant \log l(x) \leqslant 50$. Therefore Proposition 3.1.28 can be said to show that $P_{k_c(x)}$ with $m = Q_K$ defines a probabilistic sufficient statistic that contains only Halting information. By Proposition 3.2.1, any SS contains almost all information in this $P_{k_c(x)}$. Therefore, $P_{k_c(x)}$ is approximately the minimal sufficient statistic. Remind that in this Chapter, m is assumed to be a fixed universal semimeasure, with $K(m) \leqslant_+ 0$.

Also, remind the definition of *bb* as the inverse prefix-free Buzzy Beaver function in Subsection 3.1.3.

Proposition 3.2.1. *There is a c' such that for any c, if P is a probabilistic c-SS for x, then one has:*

$$K(P_{k_{c+c'}(x)}|P^*) \leqslant_+ bb(x) + 2\log k_{c+c'}(x).$$

Proof. Using P^* and a shortest program for an upper bound u of x, all values $-\log P(y)$ can be evaluated, and a Shannon-Fano code can be encoded and decoded for $y \leqslant u$. Remark that there among these codes, there is a code $E(x)$ such that $K(P) + l(E(x)) \leqslant K(x) + c + 1$. Let s be the maximal computation time to needed to decode all these $y \leqslant u$. One has that $K_s(x) \leqslant_+ K(P) - \log P(x) + c$. This shows that for some constant c' one has that $s \geqslant t_{k_{c+c'}}$. This shows that $s, k_{c+c'}(x) \longrightarrow P_{k_{c+c'}(x)}$. Therefore $P_{k_{c+c'}(x)}$ is computed by $u, P^*, k_{c+c'}(x)$. □

3.2.2 A minimal sufficient statistic can carry non-Halting information

Proposition 3.2.1 shows that a $2c$-SS can carry amount of initial bits of the Halting probability in some form. The question may be raised if at least symbolically a probabilistic minimal sufficient statistic can carry a substantial amount of non-Halting information. The answer is positive by Proposition 3.2.2.

Proposition 3.2.2.

$$\begin{aligned} \forall c, e \exists^\infty x \qquad & \left[l_e^Z(x) \geqslant_+ (I(x;H))^c\right] \\ \forall e \exists \nu > 0 \exists^\infty x \qquad & \left[l_e^Z(x) \geqslant_+ \nu l(x) + I(x;H)\right]. \end{aligned}$$

The proof is long and technical, therefore a sketch of the proof is given first. Let x^* be a program of length $K(x)$ that produces x. If P is a probabilistic SS, then it will be shown that

$$x^* \longrightarrow_+ P, K(P).$$

This means that a shortest program for x generates $K(P)$. If P where equivalent with $\Omega^{n,i}$ for some i, then i can be computed from $x, K(x)$. However, an x will

be constructed such that x^* has a computational m-sophistication of i, but i has a high complexity given x^*. At the same time, it is guaranteed that there are only bits of Ω^i_m in x. This shows that x^* does not compute i, and that there can be no SS P of length i. This also shows that the logarithmic terms in Proposition 3.1.28 are necessary. Since i has a large complexity given x^*, also numbers close to i have large complexity given x^*. This will allow to derive lower bounds for the minimal sufficient statistic relative to the m-sophistication.

Proposition 3.2.2 is proved as a corollary of Proposition 3.2.9, which shows a similar result with length conditional quantities. Before Proposition 3.2.9 is proved, Lemmas 3.2.3-3.2.8 are proved. Let m be fixed for this subsection, such that $K(m) \leqslant_+ 0$.

Lemma 3.2.3. *Let $x \in 2^n$, and $i \leqslant n/2$ such that*

$$\begin{aligned} K(x|i^*,n) &=_+ n \\ x_{i+1} &= 1. \end{aligned}$$

There is an $y \in 2^{n/2}$ such that:

$$\begin{aligned} x^i &= y^i \\ y &< x^{n/2} \\ K(y|n) &=_+ n/2 \\ K(i|y) &=_+ K(i|n) \\ I(y;\Omega_m) &\leqslant_+ i. \end{aligned}$$

Proof. Applying additivity of prefix-free Kolmogorov complexity:

$$\begin{aligned} K(x^i|i^*,n) &=_+ K(x|i^*,n) - K(x_{i\ldots n}|(x^i)^*,i^*,n) \\ &\geqslant_+ n-(n-i) \\ &\geqslant_+ i, \end{aligned}$$

and therefore: $K(x^i|i^*,n) =_+ i$. Remark that $x^i \longleftrightarrow (x^i)^*$.
Choose $v \in 2^{n/2-i-1}$ such that

$$K^H(v|x^i,i^*,n) \geqslant n/2-i-1.$$

Such v always exists. Let $y = x^i 0 v$. Obviously, the first two conditions of the Lemma are satisfied.
Applying additivity of prefix-free Kolmogorov complexity:

$$\begin{aligned} K(y|i^*,n) &=_+ K(x^i|i^*,n) + K(v|x^i,i^*,n) \\ &=_+ i + n/2 - i - 1 \\ &=_+ n/2. \end{aligned}$$

Therefore, also the third condition is satisfied.

Remark that $y \longleftrightarrow y^*$ such that:

$$\begin{array}{lll} K(i|y) & =_+ & K(i,y|n) - K(y|n) \\ & =_+ & K(y|i,n) + K(i|n) - K(y|n) \\ & \geqslant_+ & n/2 + K(i|n) - n/2 \\ & = & K(i|n). \end{array}$$

Therefore, also the forth condition is satisfied.
Remark that:

$$\begin{array}{rll} K^H(y|n) & \geqslant_+ & K^H(v|n) \geqslant_+ n/2 - i \\ K(y|n) & \leqslant_+ & n/2. \end{array}$$

Therefore, also the fifth condition is satisfied. □

Lemmas 3.2.4 and 3.2.5 show that if i can not be computed from x, then also numbers in some neighborhood of i can not be computed from x. Let $\log^{(k)} i$ be the k-th iteration $\log ... \log i$.

Lemma 3.2.4. *Let c be constant, if*

$$K(i|x) \geqslant_+ \log i + \log^{(2)} i + \log^{(3)} i,$$

then

$$\min\{K(j|x) : i^{1/c} \leqslant j \leqslant i^c\} \geqslant \log^{(3)} i - O(\log^{(4)} i).$$

Proof. The proof of the conditioned version on x is the same as the unconditioned version, which will be shown here.

$$\begin{array}{lll} K(i|n) & =_+ & K(i, \log i, \log^{(2)} i, \log^{(3)} i|n) \\ & =_+ & K(i|(\log i)^*, (\log^{(2)} i)^*, (\log^{(3)} i)^*|n) \\ & & +K(\log i|(\log^{(2)} i)^*, (\log^{(3)} i)^*, n) \\ & & +K(\log^{(2)} i|(\log^{(3)} i)^*, n) \\ & & +K(\log^{(3)} i, n). \end{array} \tag{3.2}$$

Since $K(w|\log w) \leqslant_+ \log w$ and $K(w|n) \leqslant_+ 2\log w$, we have that

$$\begin{array}{lll} K(\log^{(2)} i|n) & \geqslant_+ & K(\log^{(2)} i|(\log^{(3)} i)^*|n) \\ & =_+ & K(i|n) - K(i|(\log i)^*, (\log^{(2)} i)^*, (\log^{(3)} i)^*|n) \\ & & -K(\log i|(\log^{(2)} i)^*, (\log^{(3)} i)^*, n) - K(\log^{(3)} i|n). \\ & \geqslant_+ & \log^{(3)} i - O(\log^{(4)} i). \end{array}$$

Remark that:

$$\log^{(2)} i =_+ \log(1/c \log i) \leqslant \log^2 j \leqslant \log(c \log i) =_+ \log^{(2)} i.$$

therefore,

$$K(j|n) \geqslant_+ K(\log^{(2)} j) \geqslant_+ \log^{(3)} i - O(\log^{(4)} i).$$

□

Lemma 3.2.5. *For any c, let $\tilde{i}$ be the c most significant bits of i. If $i(1-2^{-c}) \leqslant j \leqslant i(1+2^{-c})$, then $K(j|n) \geqslant_+ K(\tilde{i}|n)$.*

Proof. Trivial. □

In the proof of Proposition 3.2.2 an i will be needed that both satisfies the conditions of Lemmas 3.2.3 and 3.2.4.

Lemma 3.2.6. *For any n and $x \in 2^n$ such that $K(x|n) =_+ n$, there is an $i \leqslant \frac{1}{2}n$ such that*

$$\begin{aligned} K(i|x,n) &\geqslant_+ \log i + \log^{(2)} i + \log^{(3)} i \\ K(x|i^*,n) &\geqslant_+ n \\ x_i &= 1. \end{aligned}$$

Proof. Let $a_2 \in 2^{\log^{(3)} n - 1}$ such that $K(a_2|x,n) =_+ \log^{(3)}(n)$,
let $a_1 \in 2^{a_2}$ such that $K(a_1|a_2^*,x,n) =_+ a_2$, and
let $i \in 2^{a_1}$ such that $K(i|a_1^*,a_2^*,x,n) =_+ a_2$,
where a_1^* and a_2^* are defined such that additivity of Kolmogorov complexity in Equation (3.3) can be applied. Remark now that:

$$\begin{aligned} a_1 &\in 2^{<\log^{(2)} n} \\ i &\in 2^{<\log n} \end{aligned}$$

When i is considered as an element of ω, one has

$$i \leqslant \frac{1}{2}n.$$

By additivity one has

$$\begin{aligned} K(i|x,n) &=_+ K(i,a_1,a_2,x,n) \\ &=_+ K(i|a_1^*,a_2^*,x,n) + K(a_1|a_2^*,x,n) + K(a_2|x,n) \quad (3.3) \\ &=_+ \log i + \log^{(2)} i + \log^{(3)} n. \end{aligned}$$

This shows the first inequality of the lemma.

From Equation (3.3) it also follows that

$$\begin{aligned} K(i|x,n) &=_+ a_1 + a_2 + \log^{(3)} n \\ &=_+ K(i|a_1,a_2,n) + K(a_1|a_2,n) + K(a_2|n) \\ &\geqslant_+ K(i,a_1,a_2|n) =_+ K(i|n). \end{aligned}$$

Remind that $x \longleftrightarrow x^*$. By additivity one has

$$\begin{aligned} K(x|i^*,n) &=_+ K(x,i|n) - K(i|n) \\ &=_+ K(x,i|n) - K(i|x,n) \\ &=_+ K(x|n) \geqslant_+ n. \end{aligned}$$

This shows the second inequality of the lemma.

Remark that from this inequality, it follows as in the proof of Lemma 3.2.3 that $K(x^i|i^*,n) \geqslant_+ i$. Suppose that for some c one has that $x_{i-c} = \cdots = x_i = 0$, then it would follow that $K(x^i|i^*,n) \leqslant_+ i - c + 2\log c$, which implies $c \leqslant_+ 0$. Therefore, for some $c \leqslant_+ 0$, $x_{i-c} = 1$. Choosing $i \leftarrow i - c$ shows the third condition of the lemma. □

Definition 3.2.7. A *length conditional semimeasure* $P(x|n)$ is a Real function from $2^{<\omega}$ into $[0,1]$ such that for all n

$$\sum\{P(x|n) : x \in 2^n\} \leqslant 1.$$

Lemma 3.2.8. *For any length conditional semimeasures $P, Q \in \dot{\Sigma}$ such that there exists a c with abs $\lim_t \log P_t/Q_t \leqslant c$, there exists a constant c' such that for t_i sufficiently above n one has*

$$\log P_{t_i}(x|n) \geqslant \log Q_{t_{i-2\log i-c'}}(x|n) - c$$

Proof. Remark that

$$f(t,n) = \min\{s : \forall x \in 2^n \Big[\log \frac{P_s(x|n)}{Q_t(x|n)} \geqslant -c \Big]\}$$

is computable, and satisfies for all t, n and $x \in 2^n$:

$$\log P_{f(t,n)}(x|n) \geqslant \log Q_t(x|n) - c.$$

Applying the length conditional variant of Lemma 3.1.7 to show that for any computable function $f(t,n)$ one has for some large enough constant e, t sufficiently larger than n:

$$t_i \geqslant BB(i-e) \geqslant f(BB(i-2e),n) \geqslant f(t_{i-2\log i-c'},n).$$

□

In a similar way as probabilistic SS, a *length conditional probabilistic SS* of an $x \in 2^n$ is a length conditional semimeasure P such that

$$K(P|n) - \log P(x|n) =_+ K(x|n).$$

Let $l^P_{x|n}$ be the length of the minimal length conditional probabilistic SS. Let $m(x|n)$ denote a universal length conditional semimeasure. The *length conditional $m(x|n)$-sophistication* is given by the following equations:

$$\begin{aligned} \Omega_{m|n,t} &= \sum\{m_t(x|n) : x \in 2^n\} \\ t_{k|n} &= \min\{t : \Omega_{m|n} - \Omega_{m|n,t} \leqslant 2^{-k}\} \\ k_c(x|n) &= \min\{k : K_{t_{k|n}}(x|n) = K(x|n)\}. \end{aligned}$$

Proposition 3.2.9.

$$\begin{aligned} \forall c, n \exists x \in 2^n \quad & \big[l^P_{x|n} \geqslant_+ (k_e(x|n))^c \\ & \wedge I(x; H|n) \leqslant_+ k_e(x|n) + 2\log k_e(x|n)\big] \\ \exists \nu > 0 \forall n \exists x \in 2^n \quad & \big[l^P_{x|n} \geqslant_+ (k_e(x|n))^c \\ & \wedge I(x; H|n) \leqslant_+ k_e(x|n) + 2\log k_e(x|n)\big]. \end{aligned}$$

Proof. Only the first inequality is shown, since the argument for the second inequality is similar by replacing Lemma 3.2.4 with Lemma 3.2.5, and making some numurous, but straightforward adaptations.

First a series of times $t_1, ...t_e$ will be constructed from the Natural numbers $k \leqslant 2^{n/2}$ which defines approximations of $\Omega^{n/2}_{m|n} = \Omega^{n/2}_{m|n,t_e}$. This will lead to a construction of the mappings:

$$k \longleftrightarrow_+ z_k,$$

from the strings z_k, strings of fixed m-depth will be constructed.

Let $m_t \in \dot{\Sigma}$ be an enumeration of m, such that that for all t there is maximally one $x \in 2^n$ with

$$m_t(x|n) \neq m_{t+1}(x|n).$$

Additionally assume that for all $k < 2^{n/2}$, for witch there is a t such that

$$\Omega_{m|n,t} < k2^{-n/2} \leqslant \Omega_{m|n,t+1} \tag{3.4}$$

there is a $z_k \in 2^n$, such that $m_t(z_k|n) \leqslant 2^{-n}$ and $m_{t+1}(z_k|n) > 2^{-n}$. Remark that from each universal length conditional Σ-semimeasure such a semimeasure can be constructed. Also, remark that for any such k

$$z_k \longleftrightarrow t_k \longleftrightarrow_+ k.$$

By Lemma 3.1.2 one has $K(\Omega^n_{m|n}|n) \geqslant_+ n$. For n large enough, let $y \in 2^{n/2}$ as in Lemma 3.2.3 with $x = \Omega^n_{m|n}$, and i chosen such that the conditions of Lemmas 3.2.3 and 3.2.4 are satisfied:

$$K(i|n) \geqslant_+ \log i + \log^{(2)} i + \log^{(3)} i \tag{3.5}$$

$$K(x|i^*, n) \geqslant_+ n \tag{3.6}$$

$$x_i = 0. \tag{3.7}$$

Remark that by Lemma 3.2.6 this is possible. Also, remark that $x_i = (\Omega_{m|n})_i = 1$ and $y_i = 0$, and therefore $\Omega^{n,i-1} \leqslant y < \Omega^{n,i}$. This shows that y determines a $k \leqslant 2^{n/2}$ such that equation (3.4) is satisfied, and the corresponding t satisfies $t_{i-1} \leqslant t \leqslant t_i$. Let $z = z_k$. Remark that

$$z \longleftrightarrow y.$$

This implies that on has $K_{t_{i+O(1)|n}}(z|n) \leqslant_+ n/2$, and thus $i \leqslant_+ k_z$. At time t_i, one has $m_t(z|n) \geqslant 2^{-n}$, and by choosing $P_t = m_t$ and $Q_t = 2^{-K_t(x|n)}$ in Lemma 3.2.8, it follows that $K_{t_{i-2\log i-O(1)|n}}(z) \geqslant_+ n$. Therefore, one has $i - 2\log i \leqslant k_e(z|n)$ By Lemma 3.2.3, $I(y;H) \leqslant_+ i \leqslant_+ k_e(z|n) + 2\log k_e(z|n)$. Therefore, z satisfies the right condition of the first claim of Proposition 3.2.2.

Let P be a minimal SS. Using Shannon-Fano code, P is part of a code for x of length below $K(x|n) + O(1)$. It follows by Theorem 2.4.7 that:

$$z^* \longrightarrow_+ P \longrightarrow_+ l_z^P. \tag{3.8}$$

By Lemma 3.2.4 it follows that for any j with $i^{1/c} \geqslant j \geqslant i^c$:

$$K(j|z) \geqslant \log^{(3)} i - O(\log^{(4)}),$$

and therefore, assuming $\log^{(3)} i > O(1)$ one has:

$$z^* \not\longrightarrow_+ j. \tag{3.9}$$

Combined with equation (3.8), this shows that either $l_z^Z < i^{1/c}$ or either $l_z^Z > i^c$. By Proposition 3.2.1 it follows that

$$l_z^Z \geqslant_+ k_e(z|n) - 2\log k_e(z|n) \geqslant_+ i - 2\log i,$$

and therefore, $l_z^Z > i^c$. This shows the left condition of the first claim of Proposition 3.2.2. □

Proof of Proposition 3.2.2. As in the proof of Lemma 3.1.14, there is a computable function f such that for infinitely many n one has $K_{f(n)}(n) = K(n)$. For such n, and $x \in 2^n$, one has $K_g(t)(x|n) + K(n) \leqslant_+ K_t(x)$, showing that $k_{c+c'}(x|n) \geqslant k_c(x)$. On the other side, for the constructed x, one has $k_c(x|n) =_+ k_{c+h(n)}$ for some small monotone non-decreasing function h. This shows that for the constructed x one has

$$k_c(x|n) =_+ k_{c+c'}(x).$$

□

Definition 3.2.10. A string x has *c-stable sophistication* iff

$$k_0(x|n) \leqslant k_c(x|n) + c.$$

Corollary 3.2.11. *For some c large enough, and for x as constructed in the proof of Proposition 3.2.9, x has c-stable sophistication.*

Proof. This follows since $x \longleftrightarrow_+ y$, and y is incompressible. □

Proposition 3.2.2 shows that there can be a difference between the minimal SS and the information carried in the initial bits of the Halting sequence. However, the proposition does not address the question whether this difference is substantial with respect to an attempt to interpret algorithms that were designed inspired by the use of minimal SS [16]. The first claim of Proposition 3.2.2 can only be satisfied for n sufficiently large, compared to the $O(1)$ constants. To obtain equation (3.9) it is assumed that $\log^{(3)} i \geqslant O(1)$, therefore,

$$n > i > 2^{2^{2^{O(1)}}}.$$

Even if it is assumed that the arbitrary constants are very low, suppose that $O(1) = 4$ could be chosen in the above equation, the corresponding n is much larger than the length of any data that can possibly be the input of an algorithm. In the proof of the second equation of Proposition 3.2.2, the constructed ν satisfies $\nu \leqslant 2^{-c}$, which implies that for large c the largest fraction of the information of the minimal SS of the constructed z in the proof is Halting information.

3.2.3 Set sufficient statistics

The question can be raised whether a set variant of the probabilistic sufficient statistic (probabilistic SS) can have a different behavior than probabilistic SS regarding the questions addressed in this section, and other questions. Proposition 3.2.13 shows that the class of probabilistic and set SS of an x only differ with respect to the information contained in $K(x)$. This shows that the results in this section can be reformulated with set SS. However, such a formulation is not so elegant but the conclusions remain the same.

Definition 3.2.12. For any function f, a finite set S is an *set f-sufficient statistic* for x iff

$$K(S) + \log |S| \leqslant K(x) + f(x).$$

Proposition 3.2.13. *There is a constant c such that for all x*

- *and for any function f, every set f-sufficient statistic S for x computes a probabilistic $f + c$ sufficient statistic P for x:*

$$S^* \longrightarrow_+ P^*, K(x).$$

- *every probabilistic c'-sufficient statistic x computes a set $c' + c$-sufficient statistic S for x with*

$$P^*, K(x) \longrightarrow_+ S.$$

Proof. The first claim of the proposition has the same proof as in [71], and it is easily observed from the definition of a minimal sufficient statistic S of x that $S \longrightarrow K(x)$. The second claim is now shown. Let

$$S_{P,k} = \{y : \text{abs}(-\log P(y) - k) \leqslant c'\}.$$

Remark that $P, k \longrightarrow S_{P,k}$, and

$$P(S_{P,k}) = |S_{P,k}|2^{-k}.$$

For all $k = ic'$, for $i \in \omega$, the sets $S_{P,k}$ are disjunct. Therefore, one can define a semimeasure for such k:

$$Q(k) = P(S_{P,k}).$$

By the Coding Theorem this shows that

$$K(k|P^*) \leqslant_+ -\log Q(k) \leqslant_+ -\log P(s_{P,k}) \leqslant_+ -\log |S_{P,k}| + k.$$

Remark that also for general k one has $K(k|P^*) =_+ -\log |S_{P,k}| + k$. For $k = -\log P(x)$:

$$\begin{aligned} K(x) &\geqslant_+ K(P) + k \\ &\geqslant_+ K(P) + K(k|P^*) + \log |S_{P,k}| \\ &\geqslant_+ K(S_{P,k}) + \log |S_{P,k}|. \end{aligned}$$

Remark that

$$S_{P,k} \longrightarrow k, P^* \longrightarrow k + l(P^*) = K(x).$$

□

3.3 Weak sufficient statistics and typical models

In this section it is always assumed that $x \in 2^n$, and that all oupi-interpreters and other computing devices have access to (an oracle for) n. By this, one has $K(n) =_+ 0$, and for any $y \in 2^{<\omega}$: $K(y|n) =_+ K(y)$.

A criterion is provided for which the WSS is equivalent with an SS, and it is shown that this criterion is always satisfied within very small error. An explicit construction will be given to convert an initial segment of the Halting sequence into a minimal WSS, and to convert a minimal WSS into an initial segment of the Halting sequence. Typical models are introduced, which have a little more natural definition than WSS'es. It is shown that WSS define typical models. For x with "stable" sophistication, the explicit constructed for a WSS is a minimal typical model, and thus also a minimal typical WSS. Finally, the question of equivalence between minimal sufficient statistics, and typical models is generalized to some mathematically simple, but non-trivial series of questions, which will be also referred to in Chapter 6.

3.3.1 Weak sufficient statistics

The reason why a minimal SS, as defined previously, is not equivalent with an initial segment of the Halting probability, is that the length of that segment carries information that would be available in the description of x, while this information does not contribute to the compression of x. If the minimal SS is encoded such that the information of the length of the minimal SS does not "count", then there is an equivalence. It turns out that this is possible by conditioning the complexities of x, Z on $C(Z)$ in the definition of a SS, where $C(Z)$ is the Kolmogorov complexity with respect to a plain Turing machine. Let ψ be a fixed universal plain Turing Machine, then $C(x) = \min\{l(p) : \psi(p, n) \downarrow = x\}$. The following equation relates prefix-free and plain Kolmogorov complexity [25, 44]:

$$C(x) =_+ K(x|C(x)) \tag{3.10}$$

Definition 3.3.1. Let $x \in 2^n$.

- A finite set $S \subset 2^n$ is a *weak f-sufficient set statistic* of a binary string x iff $x \in S$, and
$$C(S) + \log |S| \leqslant K(x|C(S)) + f(x). \tag{3.11}$$
- A computable probability distribution P over 2^n is a *weak f-sufficient probabilistic statistic* of a binary string x iff
$$C(P) - \log P(x) \leqslant K(x|C(P)) + f(x). \tag{3.12}$$
- A computable prefix-free function $F : 2^{<n} \to 2^n$ is a *weak sufficient function statistic* of a binary string x iff
$$C(F) - \log P(x) \leqslant K(x|C(F)) + f(x).$$

For $Z = S, P, F$, the *minimal weak sufficient statistic* Z'_x is the weak sufficient statistic Z such that $C(Z)$ is minimal within some constant. Let $l'^Z_x = C(Z'_x)$.

In the same way as in Lemma 3.1.25 for any x, a probabilistic weak sufficient statistic (probabilistic WSS) is algorithmically equivalent with a function WSS. In the same way as in Proposition 3.2.13, a set WSS S is equivalent with a probabilistic WSS combined with some description for $K(x|C(S))$.

Let $||\log Z||$ be either $\log |S|, -\log P(x)$, or $\min\{l(d) : F(d) = x$. Then the defining equation for a f-WSS is given by:

$$C(Z) + ||\log Z|| \leqslant K(x|C(Z)) + f(x).$$

which contrasts with the generic definition of f-SS:

$$K(Z) + ||\log Z|| \leqslant K(x) + f(x).$$

Propositions 3.3.2 and 3.3.3 shows that the amount of c-WSS'es for a fixed x is much larger than the amount of SS'es.

Proposition 3.3.2. *For every c, there is an upper bound on the amount of different c-SS'es an $x \in 2^{<\omega}$ can have.*

Proof. This follows by Theorem 2.4.7 and the definition of SS. □

Proposition 3.3.3. *For c large enough, x has at least $K(x) - k_0(x)$ different c-WSS'es, where k_0 is defined for $m = Q_p$.*

Proof. All programs Halting on ϕ can be enumerated $p_0, p_1, \ldots$ in some non-decreasing order of Halting times $t[p_0], t[p_1], \ldots$ Let P be the semimeasure $P(i) = 2^{-l(p_i)}$, and let α_i be the corresponding Shannon-Fano code for i. Remark that $\alpha_0 < \alpha_1 < \ldots$. Let

$$S(i,j) = \{k : \phi(p_k) \downarrow = x \wedge \alpha_i^j = \alpha_k^j\}$$

If for some j one has

$$\alpha_i^j + 2^{-j} \leqslant \Omega_{Q_p}, \tag{3.13}$$

then $\alpha_i^j \longrightarrow \alpha_i^j$. Let

$$Q_{i,j}(x) = N \sum \{2^{j - l(\alpha_k)} : k \in S(i,j)\},$$

where N is a normalization constant such that Q defines a measure. Remark that if i, j satisfies Equation (3.13), then $\alpha_i^j \longrightarrow Q_{i,j}$. For any i such that p_i is a witness of $K(x)$, and for $j \geqslant k_0(x)$, one has that for $m = Q_p$, equation (3.13) is satisfied, and thus for for all

$$k_0(x) \leqslant j \leqslant l(\alpha_i) \leqslant K(x) + 1.$$

One has that $Q_{i,j}(x)$ defines a measure computable from α_i^j. This implies that

$$C(Q_{i,j}) - \log Q_{i,j}(x) \leqslant_+ K(x),$$

and therefore, $Q_{i,j}$ is a probabilistic WSS. □

The question raises whether every weak sufficient statistic is also a sufficient statistic. The proposition below gives a condition.

Proposition 3.3.4. *For $Z = S, P, F$, for $c \leqslant_+ 0$, and for some c' large enough, if Z is a c-SS of $x \in 2^n$, and*

$$Z, K(Z) \longrightarrow_+ C(Z),$$

then Z is a $c + c'$-WSS of x.

Proof. Remark that by equation (3.10) every c-WSS Z defines a shortest description of x given $C(Z)$ on a prefix-free Turing machine. Remark that Z defines a shortest program for x given $C(Z)$. By the conditioned version of Theorem 2.4.7, it follows that

$$x, K(x|C(Z)), C(Z) \longrightarrow_+ Z.$$

By the assumption of the proposition

$$Z^* \longleftrightarrow_+ C(Z).$$

One also has $K(x) = K(x, K(x))$, and its conditioned equivalent. Therefore:

$$\begin{aligned} K(x|C(Z)) &=_+ K(x, K(x|C(Z))|C(Z)) \\ &=_+ K(x, Z|C(Z)) \\ &=_+ K(x|Z^*, C(Z)) + K(Z|C(Z)) \\ &=_+ K(x|Z^*) + K(Z|C(Z)) \\ &=_+ K(x) - K(Z) + K(Z|C(Z)). \end{aligned}$$

$$||\log Z|| =_+ K(x) - K(Z) =_+ K(x|C(Z)) - C(Z).$$

□

The question raises whether $Z, K(Z) \longrightarrow_+ C(Z)$. Let ${}^k2$ be the tetration with base 2 and height k, it is the k-th iteration of taking the power of 2:

$$2^{(2^{(\ldots^2)})}.$$

The inverse of the tetration function is the *super-logarithm*, it is

$$\text{slog}x = \max\{k : {}^k2 \leqslant x\}.$$

Lemma 3.3.5.

$$K(C(x)|x, K(x)) \leqslant_+ O(slogx).$$

Proof. $C(x)$ is approximated as:

$$\begin{array}{rclcl} k_1 & = & K(x) & & \\ k_2 & = & K(x|k_1^*) & = & K(x|K(x)^*) \\ k_3 & = & K(x|k_2^*) & = & K(x|K(x|K(x)^*)^*) \\ k_i & = & K(x|k_{i-1}^*) & = & K(x|K(x|...^*)^*). \end{array}$$

Remark that since $k_1 \leqslant_+ 2\log x$, it follows that $k_1 - k_2 \leqslant_+ 2\log^{(2)} x$. Suppose that

$$\text{abs}(k_{i-1} - k_i) \leqslant_+ 2\log^{(i)} x,$$

then it follows that

$$\begin{aligned} \text{abs}(k_i - k_{i+1}) &\leqslant_+ \text{abs}(K(x|k_i^*) - K(x|k_{i+1}^*)) \\ &\leqslant_+ 2\log \text{abs}(k_i - k_{i+1}) \\ &\leqslant_+ 2\log^{(i+1)} x. \end{aligned}$$

and therefore the series has converged after slog$x$ steps, within a constant. The limit of the series is some k for which $K(x|k^*) =_+ k$. There is only one value k that for some x satisfies $K(x|k^*) =_+ k$. Since if there was also a $l < k$ such that $K(x|l^*) =_+ l$, then

$$k - l =_+ K(x|k^*) - K(x|l^*) \leqslant_+ 2\log(k - l),$$

and therefore, $k =_+ l$. Remark that the proof of equation (3.10), see [44, Lemma 3.1.1] also shows that

$$C(x) =_+ K(x|C(x)^*).$$

Therefore, it follows that this series k_i converges to $C(x)$. To prove the proposition, it remains to show that

$$x, k_i, K(x, k_i) \longrightarrow_+ k_{i+1}, K(x, k_{i+1})$$

Remark that by 2.4.8 that

$$x, k_i, K(x, k_i) \longrightarrow_+ k_{i+1}.$$

In the same way

$$K(x, k_i, k_{i+1}) \longrightarrow_+ K(x, k_{i+1}).$$

□

By Lemma 3.3.5 and Proposition 3.3.4, it can be stated that for strings of realistic length, every WSS is a SS. This is why the name *weak* sufficient statistic was chosen. It contrasts with the name *strong* sufficient statistic defined in [49].

3.3.2 Explicit weak sufficient statistics

An explicit construction of a probabilistic WSS P'_x for an $x \in 2^n$ is now given. Remark that in [26] a construction is given of what is called an "Explicit minimal near-sufficient statistic". The construction there can be adapted to a construction of a set MWSS using the same ideas as as the construction of P'_x. The construction of P'_x makes use of k'_x, a variation of m-sophistication, which will be called BB-sophistication since it uses the Buzzy Beaver function. Let ψ be an universal optimal plain Turing machine.

Definition 3.3.6.

$$\begin{aligned} BB(k) &= \max\{\psi(p) : p \in 2^k\} \\ k_c^{BB}(x) &= \min\{k : K_{BB(k)}(x|k) \leqslant K(x|k) + c\} \end{aligned}$$

The following Lemma shows a relation between conditional m-sophistication and BB-sophistication, which reminds to equation (3.10).

Lemma 3.3.7.

$$k_c^{BB}(x) =_+ k_c(x|k_c^{BB}(x))$$

Proof. It suffices to show that:

$$\begin{aligned} BB(k) &\leqslant t_{k+O(1)|k} \\ t_{k|k} &\leqslant BB(k + O(1)) \end{aligned}$$

The first inequality is now shown. Remark that any program of length k halting on a plain Turing machine, can be adapted to a program of length $k + O(1)$ by adding a constant amount of instructions, halting on a prefix-free Turing machine given k. By subsequently applying the conditional variant of Lemma 3.1.7, the equation is shown.

The second inequality follows by remarking that $\Omega^k_{m|k}$, the conditional version of Ω^k_m, defines a Halting program on plain Turing machine that outputs $t_{k|k}$ by adding a finite amount of instructions. □

For some e large enough let $k = k_0^{BB}(x)$ in

$$P'_x(y) = \begin{cases} 2^{-K_{BB(k)}(y|k)+k-e} & \text{if } K_{BB(k}(y|k) \neq K_{BB(k-1)}(y|k), \\ 0 & \text{otherwise.} \end{cases}.$$

Proposition 3.3.8. *For c, e large enough P'_x is a probabilistic c-WSS.*

Proof. First it is shown that P'_x is a semimeasure. Let m be the universal length conditional semimeasure given by

$$m_t(y|l, n) = 2^{-K_t(y|l)}.$$

For e large enough, and for any k, by Lemma 3.3.7:

$$\begin{aligned} &\sum_y m_{BB(k)}(y|k, n) - m_{BB(k-1)}(y|k, n) \\ &\leqslant \sum_y m_{t_{k+e|k}}(y|k, n) - m_{t_{k-e|k}}(y|k, n) \\ &\leqslant \Omega_{m|k} - \Omega_{m|k,t_{k-e|k}} \leqslant 2^{-k+e}. \end{aligned}$$

Choosing e in the definition of P'_x large enough, shows that for all c: P'_x is a semimeasure.

It remains to show that P'_x satisfies the defining equation (3.12) of a probabilistic $c + c'$-WSS. Remark that given $C(P)$, a program for P on a plain Turing machine can be turned into a program for P given $C(P)$ on a prefix-free Turing machine by adding a constant amount of instructions. Therefore $C(P) =_+ k_0^{BB}(x)$. Using Shannon-Fano code, this shows that

$$C(P) - \log P(x) \geqslant_+ K(x|C(P)).$$

Remark that

$$C(P'_x) \leqslant_+ k_0^{BB}(x).$$

By the choice of m, one also has for $k = k_0^{BB}(x)$ that

$$P'_x(x) = m_{BB(k)}(x|k,n) - m_{BB(k-1)}(x|k,n) \geqslant 1/2 m_{BB(k)}(x|k,n) = 2^{-K(x)+k-e-1}.$$

Combining the previous three equations shows inequality (3.12). □

3.3.3 Minimal typical models

Set typical models (set TM) were studied in [70], and is shown that within logarithmic bounds, the complexity of an almost minimal set TM and an almost minimal SS are related. It is shown here that a minimal probabilistic TM is equivalent with a minimal probabilistic WSS, and therefore, there complexities are equal within constant bounds. It is shown that a minimal TM some initial segment of the Halting .

Definition 3.3.9. Let $f \in \Delta_1$, and $x \in 2^{<\omega}$.

- Let S^* denote the shortest program on a plain Turing machine. A finite set S is a *set f-typical model* for x iff $x \in S$ and

$$\log |S| \leqslant K(x|S^*) + f(x).$$

- Let P^* denote the shortest program on a plain Turing machine that computes P. A computable semimeasure P is a *probabilistic f-typical model* for x iff

$$-\log P(x) \leqslant K(x|P^*) + f(x).$$

- Let F^* denote the shortest program on a plain Turing machine that computes F. A computable function $F : \omega \to \omega$ is a *function f-typical model* for x iff

$$\exists d \Big[F(d) = x \wedge l(d) \leqslant K(x|F^*) + f(x) \Big].$$

For $Z = S, P, F$, a *minimal typical model* is a typical model Z such that $K(Z)$ is minimal within a constant.

The same proof of Proposition 3.2.13 also shows that the function typical models are equivalent with the probabilistic typical models. Also, a similar equivalence as Proposition 3.2.13 can be shown. Remark that in [70], a set typical model is defined as $\log |S| =_+ K(x|S)$. In this definition S is replaced by its minimal description, with respect to the plain Turing machine ϕ. Since [70] only considers equalities of functions within $O(\log l(x))$-terms, both in value and in argument. The results shown there, also remain valid using the definition above. By Lemma 3.3.5, the results also hold within $O(\text{slog})$ terms, if Z^* was the shortest representation on a prefix-free Turing machine.

Proposition 3.3.10. *There exists a c such that every f-WSS for $x \in 2^n$ is also an $f + c$-typical model (TM) for $x \in 2^n$.*

Proof. Remind that for any WSS Z:

$$x, C(Z), K(x|C(Z)) \longrightarrow_+ Z.$$

Therefore:

$$\begin{aligned} K(x|Z^*) &=_+ K(x|Z^*, C(Z)) \\ &=_+ K(x, Z|C(Z)) - K(Z|C(Z)) \\ &=_+ K(x|C(Z)) - K(Z|C(Z)) \\ &=_+ ||\log Z||, \end{aligned}$$

where $||\log Z||$ is either $\log |S|, -\log P(x)$, or $\min\{l(d) : F(d) = x\}$. □

By the same example as in [70], it follows that there are TM's that are not WSS'es. According to Proposition 3.3.11, P'_x defines a minimal TM, and by Proposition 3.3.13 a minimal WSS is equivalent with the minimal TM, which is equivalent with an initial segment of the Halting sequence.

Proposition 3.3.11. *For any constant c, there is a constant c' such that if P is a probabilistic c-TM for x, then $C(P) \geqslant k^{BB}_{c'}(x) - c'$.*

Proof. Let "$\leqslant^+$" mean "less than within a constant possibly dependent on c", and similar for $\geqslant^+$ and $=^+$. Let P be a TM, then it will be shown that

$$K(x|C(P)) =^+ K_{BB(C(P)+O(1))}(x|C(P)). \tag{3.14}$$

If $C(P)$ was unboundedly below k'_x, this would contradict the definition of k'_x.

Therefore it remains to show equation (3.14).

$$\begin{aligned} C(P) - \log P(x) &=^+ C(P) + K(x|P^*) \\ &=^+ K(P|C(P)) + K(x|P^*, C(P)) \qquad (3.15) \\ &=^+ K(x, P|C(P)) \\ &=^+ K(x|C(P)) + K(P|x, K(x|C(P)), C(P)) \qquad (3.16) \end{aligned}$$

On the other side, let s be the computation time to compute $-\log P(z)$ for all $z \in 2^n$ from P^*. Than $K_s(x|P^*) =^+ -\log P(x)$. For computable functions f, g large enough we have by Lemma 2.4.5:

$$\begin{aligned} C(P) - \log P(x) &=^+ C(P) + K_s(x|P^*) \\ &\geqslant^+ K_{g(s)}(x, P|C(P)) \\ &\geqslant^+ K_{f(s)}(x|C(P)) + K_{f(s)}(P|x, K_{f(s)}(x|C(P)), C(P)). \end{aligned}$$

Let $\Delta = K_{f(s)}(x|C(P)) - K(x|C(P)) \geqslant 0$, then combining equations (3.16) and (3.17):

$$\begin{aligned} &K(x|C(P)) + K(P|x, K(x|C(P)), C(P)) \\ &\geqslant^+ K_{f(s)}(x|C(P)) + K_{f(s)}(P|x, K_{f(s)}(x|C(P)), C(P)) \\ &\geqslant^+ K(x|C(P)) + \Delta + K(P|x, K(x|C(P)), C(P)) - 2\log\Delta. \end{aligned}$$

This shows that $0 \geqslant^+ \Delta - 2\log\Delta$, and therefore $\Delta =^+ 0$. Since $BB(C(P) + O(1)) \geqslant s$, equation (3.14) is satisfied. □

Remind Definition 3.2.10.

Corollary 3.3.12. *For c, c' large enough, if x is c-stable, then P'_x defines probabilistic c'-TM of minimal complexity within a constant.*

Proof. By Proposition 3.3.8, $C(P'_{0,x})$ is a c''-WSS, for some constant c''. By Proposition 3.3.10, it is also an e'-TM, for some constant e'. Assume that x is c-stable for c large enough, one has

$$C(P'_x) =_+ k_0^{BB}(x) \geqslant_+ k_c^{BB}(x).$$

Therefore, P'_x is also an c'''-WSS, with $c''' \leqslant c$. For c sufficiently above e', by Proposition 3.3.11, it follows that P'_x is minimal within a constant. □

Let H'^n be the Halting sequence relative to a plain Turing machine, conditioned. It is, $H'^n_i = 1$ if $\psi(i, n) \downarrow$, and $H'^n_i = 0$ otherwise. Proposition 3.3.13 shows that a probabilistic minimal TM is equivalent with an initial segment of $H'_{|l}$.

Proposition 3.3.13. *If P is a probabilistic TM, minimal within a constant $c \leqslant_+ 0$, and P^* its minimal description on a plain Turing machine, then*

$$P^* \longleftrightarrow_+ H'^{n,2^{k_0^{BB}(x)}} \longleftrightarrow_+ (P'_x)^*.$$

Proof. Remark that by Corollary 3.3.12, we have that $C(P) =_+ k'_0(x)$. From the proof of Proposition 3.3.11 equation (3.14) actually shows that if s is the maximal to evaluate a Shannon-Fano code according to $P(y)$ for any $y \in 2^n$, then:

$$K(x|C(P)) =_+ K_s(x|C(P)),$$

This shows that $s \geqslant BB(k_0^{BB}(x) - O(1))$. Remark that s can be computed from P, therefore,

$$s \leqslant BB(C(P) + c) \leqslant BB(k_0^{BB}(x) + c'),$$

for some c, c' large enough. Let p be the program of length $k_0^{BB}(x)$ with largest output, then

$$P \longleftrightarrow_+ p \longleftrightarrow_+ H'^{n,2^{k_0^{BB}(x)}}.$$

The last $\longleftrightarrow$ follows by definition of P'_x. □

Proposition 3.3.14. *Let $L_e^Z(x)$ be the length of the minimal e-TM of type Z.*

$$\begin{aligned} \forall c, e, e' \exists^\infty x \qquad & \left[l_e^Z(x) \geqslant_+ (L_{e'}^Z(x))^c\right] \\ \forall e \exists \nu > 0 \exists^\infty x \qquad & \left[l_e^Z(x) \geqslant_+ \nu l(x) + L_{e'}^Z(x)\right]. \end{aligned}$$

Proof. This follows from Proposition 3.2.9 and from Lemma 3.3.7 by remarking that

$$I(x; H) \geqslant kbb_0(x)) - O(\log k_0(x|k_0^{BB}(x))) =_+ k_0(x) - O(\log k_0(x)).$$

□

3.3.4 Minimal set models and open questions

For a set S, and for $F \in \{\Delta_1, \Sigma, \Pi\}$, let S^{*F} denotes the shortest F-description of S. For a sequence $\mathcal{S}$ of sets, and for some $x \in 2^{<\omega}$, let

$$S_{[x]} = \text{argmin}\,\{K(S) : x \in S \in \mathcal{S}\}.$$

Question 3.3.15. *For $F \in \{\Delta_1, \Sigma\}$: Does there exists an enumerable series of finite computable sets such that*

$$K(S_{[x]}^{*\Pi}) + K(x|S_{[x]}^{*\Pi}) - K(x)$$

grows unbounded ?

For $F = \Pi$, this question can now be answered.

Proposition 3.3.16. *There exists an enumerable series of finite computable sets such that*

$$K(S^{*\Pi}_{[x]}) + K(x|S^{*\Pi}_{[x]}) - K(x)$$

grows unbounded.

Proof. Let $p_0, p_1, \ldots$ be an enumeration of programs for finite sets $U_0, U_1, \ldots$ For some c large enough, and for all i, let $\mathcal{S}$ be the series of sets (typical models)

$$S_i = \{x \in U_i : K(x|p_i) \leqslant |U_i| + c\}.$$

Remark that for all i: p_i defines a Π-description for S_i. If $x \in S_i$ for some $i \in \omega$, then x defines a set c-TM. Moreover, S^x defines a minimal c-TM for x. Remark that it suffices to show that

$$K(x, l(S^{*\Pi}_{[x]})) - K(x) =_+ K(l(S^{*\Pi}_{[x]})|x) \tag{3.17}$$

grows unbounded. Since a similar equivalence can be shown as in Proposition 3.2.13, for typical models, it suffices to show the above equation for probabilistic typical models.

Let x, y, i as in the proof of Proposition 3.2.9. Remind that $x \longleftrightarrow_+ y$, and that $y \in 2^{n/2}$ and y is incompressible, therefore representing a shortest program for x. Therefore, if P corresponds to $S_[x]$, than P defines a minimal c'-TM for some c. Remark that by Corollary 3.2.11, x has c''-stable sophistication. Therefore, by Proposition 3.3.11 it follows that $K(P) =_+ k^{BB}_{c''}(x) =_+ k^{BB}_0(x)$. Using Lemma 3.3.7 it follows that

$$0 \leqslant_+ K(P'_0(x)) - k_0(x) \leqslant_+ 2 \log k_0(x),$$

and that $i - k_0(x) \leqslant 2 \log k_0(x)$, and that

$$K(i|y) \geqslant_+ \log i + \log^{(2)} i + \log^{(3)} i.$$

Therefore, one has

$$K(K(P)|y) \geqslant_+ \log i.$$

If P is the minimal TM, than by Proposition 3.3.13, Therefore $K(P)$ is also close to which implies equation (3.17). □

Acknowledgements

The author is grateful to P. Vitanyi who raised the question on the relation between an algorithmic minimal sufficient statistics and an initial segment of a Halting sequence.

4
Sumtests

Abstract. On one side, sumtests provide some idealization of significance tests, on the other side, sumtests raise simple questions on Kolmogorov complexity of finite strings with non-trivial answers. Three questions will be asked on sumtests for computable and lower semicomputable semimeasures:

- Is there within a computability class a largest sumtest within an additive constant (universal sumtest) ?
- Which computability class contains the largest sumtests ?
- How large can sumtests in a computability class grow ?
- Can the sumtests in a computability class be characterized by Kolmogorov complexities ?

The first section addresses the first two questions in general. The following sections address all these questions in a specific case: sumtests for a universal semimeasure, and sumtests for a direct product of two universal semimeasures. It is shown that for these semimeasures, there are no unbounded computable and lower semicomputable sumtests. Therefore, the upper semicomputable sumtests are studied. A characterization in terms of Kolmogorov complexity is given, and from this, almost tight upper bounds are shown. For universal semimeasures such a bound is given by $\log l(x) + O(\log \log l(x))$, and for a product of two universal semimeasures one obtains $l(x) - O(\log l(x))$. It is shown in both cases, that there are no universal sumtests.

Definition 4.0.17. A sumtest for a semimeasure P is a function $d : \omega \to \mathbb{Z}$ such

that

$$\sum_{x\in\omega} P(x)2^{d(x)} \leqslant 1.$$

The set of all sumtests for P is denoted as S_P. A sumtest d defines a set of x'es of small P-probability, obtained by the set of x'es with high $d(x)$. If a sumtest is high, then it can be said that either a rare event has occurred, or the model underlying P is not representative for the generation of x. If for some computability class F, a sumtest $d \in S_P \cap F$ is universal, then an optimal procedure exists to state that d is typical. While the definition of typical model allows a binary decision on the typicality of some data for x, sumtests quantise the typicality. Proposition 4.1.10 relates an optimal procedure using sumtests to the definition of typical models.

4.1 Sumtests for general Δ_1 and Σ-semimeasures

4.1.1 Which computability class contains the largest sumtests ?

For any function f, let $f^+ = \max\{f, 0\}$.

Proposition 4.1.1. *For P computable, and for any Π-sumtest d for P, there is a Σ-sumtest d' such that $d' - d^+$ is unbounded.*

Proof. If $e, e' \in S_P$ then also $\frac{1}{2}(e + e') \in S_P$. Since the zero function is in S_P, without loss of generality, it can be assumed that $d \geqslant -1$, and therefore it remains to show that $d' - d$ is unbounded. Let $(x_0, t_0), (x_1, t_1), \ldots$ be a sequence such that

$$x_0 \leqslant t_0 \leqslant x_1 \leqslant t_1 \ldots$$

and

$$P(x_i)2^{d_{t_i}(x_i)} \leqslant 2^{-2i-1}.$$

Since d is upper semicomputable such a sequence can be enumerated. Let

$$d'_t(x) = \begin{cases} d_t(x_i) + i & x = x_i \text{ and } t \geqslant t_i \text{ for some } i \\ 0 & \text{otherwise.} \end{cases}$$

Remark that $d'_t(x)$ is a Σ-function, and exceeds d'. It also defines a sumtest since

$$\begin{aligned} \sum_{x\in\omega} P(x)2^{d(x)} &\leqslant \sum_{x\in\omega} P(x)2^{d_t(x_i)+i} \\ &\leqslant \sum_{i\geqslant 1} 2^{-2i}2^i \leqslant 1. \end{aligned}$$

□

We question if also a stronger form of Proposition 4.1.1 is possible ?

Question 4.1.2. *For P computable, and for any Π-sumtest d for P, is there a Σ-sumtest d' such that $d' \geqslant^+ d$, and $d' - d$ is unbounded ?*

Is Proposition 4.1.1 also valid when the Σ and Π classes are exchanged ?

Question 4.1.3. *For P computable, and for any Σ-sumtest d for P, is there a Π-sumtest d' such that $d' - d^+$ is unbounded*[1] *?*

Remark that in Question 4.1.3 it can not be additionally required that $d' \geqslant d - c$ for some c, since otherwise it would follow from Proposition 4.1.1 that there is another Σ-sumtest d'' such that $d'' - d^+$ is unbounded, which would imply that there is no universal Σ-semimeasure, contradicting Theorem 4.1.9.

Proposition 4.1.1 does not longer hold for Σ-semimeasures P, not even in weaker form. This can be observed by considering the special case of $P = m$, and applying Corollary 4.1.5 of Lemma 4.1.4. This also shows that Question 4.1.3 is answered in the positive for $P = m$.

Lemma 4.1.4 ([7]). *For any unbounded Σ-function $d : \omega \to \mathbb{Z}$ there is a computable measure P such that*

$$\sum_{x \in \omega} P(x) 2^{d(x)} = \infty. \tag{4.1}$$

Proof. Suppose that $d : \omega \to \mathbb{Z}$ is Σ and unbounded. We construct a computable measure P such that

$$\sum_{x \in \omega} P(x) = 1. \tag{4.2}$$

and (4.1) holds. The construction is in ω stages. At stage s, search for a fresh (i.e. hitherto not used in the construction) element x such that $d(x) \geqslant s$. Such x can be found effectively since d is unbounded and Σ. For this x define $P(x) = 2^{-s}$. To make sure that P is total, define $P(y) = 0$ for all $y < x$ for which $P(y)$ was not yet defined at a previous stage. End of construction.

Clearly, the P satisfies (4.1) and (4.2), since at stage s of the construction we contribute an amount of 2^{-s} to $\sum_x P(x)$ and an amount of at least 1 to $\sum_x P(x) 2^{d(x)}$. □

Corollary 4.1.5 ([7]). *Every Σ-sumtest for the universal Σ-semimeasure m is bounded.*

Proof. Suppose that d is unbounded. Let P be as in Proposition 4.1.4. Since m is universal, there is $q > 0$ with $m(x) \geqslant qP(x)$ for all x. Then $\sum_x m(x) 2^{d(x)} \geqslant \sum_x qP(x) 2^{d(x)} = \infty$, hence d is not a sum-test for m. □

[1] I expect that this question is not so difficult, but could not go into because of time. I conjecture that this is true if $\log P(x) \leqslant^+ O(l(x))$. Maybe a good master thesis topic ?

An order is a non-decreasing unbounded function in ω.

Proposition 4.1.6 ([7]). *For every Σ-semimeasure P there is a Π-order d that is a sum-test for P.*

Proof. The idea is to monitor the tails of the sum $\sum_x P(x)$, and estimate at every stage the first element x_i such that $\sum_{y \geqslant x_i} P(y) \leqslant 2^{-i}$. The x_i may grow, but eventually come to a finite limit. If we know them we can add suitable large factors $2^{d(x)}$ that satisfy $\sum_x P(x)2^{d(x)} \leqslant 1$. If x_i turned out to be wrong, we simply decrease $d(x)$, but we have to do this only finitely often. Formally the construction proceeds as follows.

Start with $x_{i,0} = i$. At stage s, when

$$\sum_{y \geqslant x_{i,s}} P_s(y) \leqslant 2^{-i}$$

let $x_{i,s+1} = x_{i,s}$, otherwise set $x_{j,s+1} = x_{j,s} + 1$ for all $j \geqslant i$. For all $x \in [x_{i,s}, x_{i+1,s})$ define

$$d_s(x) = \lfloor \log i \rfloor.$$

End of construction.

First note that $\lim_s x_{i,s} = x_i$ exists for every i since $\sum_x P(x)$ converges. Since $x_{i,s}$ is non-decreasing, $d_s(x)$ can only decrease, and since the limit exists it can do so only finitely many times.[2] Hence $d \in \Pi$, and it is unbounded since $d(x_i) = \lfloor \log i \rfloor$. Finally,

$$\begin{aligned} \sum_{x \in \omega} P(x)2^{d(x)} &\leqslant \sum_{i \in \omega} \sum_{x \in [x_i, x_{i+1})} P(x)2^{\log i} \\ &\leqslant \sum_{i \in \omega} i \sum_{x \geqslant x_i} P(x) \\ &\leqslant \sum_{i \in \omega} 2^{-i} i = 2. \end{aligned}$$

Therefore, $d(x) - 1$ defines a sumtest for P. □

Since this answers Question 4.1.3 in the positive for a universal Σ-semimeasure, one can raise the question whether the answer of Question 4.1.3 generalizes for all Σ-semimeasures P.

Question 4.1.7. *For P computable, and for any Σ-sumtest d for P, is there a Π-sumtest d' such that $d' - d^+$ is unbounded*[3] *?*

[2] Note that since $d_0(x) = \log x$, $d_s(x)$ can change at most $\log x$ times, but the number of times $x_{i,s}$ changes is not computably bounded. Hence the limit function d can in general be very slow growing, that is, be dominated by any computable order.

[3] I conjecture that this is a hard question, with a same answer as in the computable case.

We remark that the proof of Proposition 4.1.6 is no longer valid in a length conditional setting. However, for universal length conditional Σ-semimeasures, the conclusions remain by Proposition 4.2.12.

4.1.2 Is there a universal sumtest in some computability class ?

For universal semimeasures the answer is given by Proposition 4.1.8 and Theorem 4.1.9.

Proposition 4.1.8 ([7]). *Suppose that P is a computable semimeasure, then*
- there is no universal Π-sumtest for P,
- there is no universal computable sumtest for P.

Proof. The idea is similar to that of Proposition 2.2.2. Given $d \in \Pi$ such that $\sum_x P(x)2^{d(x)} \leqslant 1$, construct $d' \in \Pi$ such that for all i there is x such that $d'(x) \geqslant d(x) + i$. Given i, effectively search for x such that $P(x)2^{d(x)} < 2^{-2i}$ (which is possible since such x exist and $d \in \Pi$), so that $P(x)2^{d(x)+i} < 2^{-i}$. For this x define $d'(x) = d(x) + i$, and set $d'(y) = d(y)$ for all $y < x$ for which $d'(y)$ was not yet defined. Then

$$\sum_{x\in\omega} P(x)2^{d'(x)} \leqslant \sum_{d'(x)=d(x)} P(x)2^{d(x)} + \sum_{i\in\omega} 2^{-i} < \infty,$$

hence $d' - c$, for some c large enough, is a Π-sum-test for P not dominated by d. □

Theorem 4.1.9. *Let m be a universal Σ-semimeasure. Suppose that $P > 0$ is a computable semimeasure, a universal Σ-sumtest is given by*

$$\log \frac{m(x)}{P(x)}.$$

Proof. Suppose that there was a Σ-sumtest d for P such that $d(x) - \log \frac{m(x)}{P(x)}$ was unbounded, then the Σ-semimeasure

$$Q(x) = P(x)2^{d(x)},$$

would exceed $m(x)$ with an unbounded factor, □

Remark that the sumtest of 4.1.9 is universal in S_P. A procedure can be constructed, that is optimal for all computable P. This allows to relate the definition of sumtests to the definition of typical models in Definition 3.3.9.

Proposition 4.1.10. *Let $\mathcal{P} \subset 2^{<\omega}$ be the set of codes p on a plain universal Turing machine ψ defining computable semimeasures P_p. Let $S_{\mathcal{P}}$ be the set of two*

argument functions $d_p(x)$ on $\mathcal{P} \times \omega$, such that for each $p \in \mathcal{P}$, one has $d_p \in S_{P_p}$. Than the function

$$d_p(x) = \frac{m(x|p)}{P(x)}$$

is universal on $\mathcal{P} \times 2^{<\omega}$ in $S_{\mathcal{P}}$.

Proof. The proof goes by a similar way as the proofs of Theorems 2.2.4 and 4.1.9. □

Applying the Coding Theorem 2.4.3, it follows that

$$d_{P^*}(x) =_+ K(x|P^*) - \log P(x). \tag{4.3}$$

For general Σ-semimeasures P with $P > 0$, such a nice correspondence is not longer true.

Proposition 4.1.11 ([7]). *There exists a strictly positive Σ-semimeasure P such that there is no Σ-universal sum-test for P.*

Proof. Since the constant zero function is a sum-test for any semimeasure, a universal sum-test is bounded from below by some constant $k \in \mathbb{Z}$. So in proving that such a universal sum-test does not exist we may restrict ourselves to such functions.

Let d_i be an effective enumeration of all Σ-functions from ω to $\mathbb{Z} \cup \{\infty\}$ that are bounded from below by some (possibly negative) constant. (The latter assumption is needed to have an effectively Σ-class of functions; for the rest of the proof it is not needed.) Let $d_{i,s}$ denote the approximation of d_i. We construct a semimeasure $P \in \Sigma$ and functions $d'_i \in \Sigma$ so that for every i it holds that $d'_i - d_i$ is unbounded and

$$\sum_x P(x)2^{d_i(x)} \leqslant 1 \Longrightarrow \sum_x P(x)2^{d'_i(x)} \leqslant 1. \tag{4.4}$$

Let $\langle x, y\rangle$ be a bijective pairing function from ω^2 to ω. We assign an infinite computable domain R_i to the strategy for d_i as follows. Define

$$R_i = \big\{\langle x, i\rangle : x \in \omega\big\}$$

and

$$d'_{i,s}(x) = \begin{cases} d_{i,s}(x) + x & \text{if } x \in R_i \\ 0 & \text{otherwise.} \end{cases}$$

We construct P by defining its approximation P_s as follows. Let $P_0(x) = 2^{-2x-1}$, so that P is strictly positive. At stage s of the construction, for every $i \leqslant s$, if s is the first stage such that

$$\sum_{x<s} P_s(x)2^{d'_{i,s}(x)} > 1 \tag{4.5}$$

then define

$$P_{s+1}(x) = P_s(x)2^{d'_{i,s}(x)-d_{i,s}(x)} = P_s(x)2^x$$

for every $x \in R_i$. Note that since this can happen only once, we have that $P_s(x)$ equals either $P_0(x)$ or $P_0(x)2^x$. This ends the construction.

We check that requirements (4.4) are satisfied for every i. Suppose that $\sum_x P(x)2^{d'_i(x)} > 1$. Then (4.5) holds for some s, hence

$$\begin{aligned}
\sum_{x\in\omega} P(x)2^{d_i(x)} &\geqslant \sum_{x\notin R_i} P_s(x)2^{d_{i,s}(x)} + \sum_{x\in R_i} P_{s+1}(x)2^{d_{i,s}(x)} \\
&\geqslant \sum_{x\notin R_i} P_s(x) + \sum_{x\in R_i} P_s(x)2^{d'_{i,s}(x)-d_{i,s}(x)}2^{d_{i,s}(x)} \\
&\geqslant \sum_{x\notin R_i} P_s(x) + \sum_{x\in R_i} P_s(x)2^{d'_{i,s}(x)} \\
&= \sum_{x\in\omega} P_s(x)2^{d'_{i,s}(x)} > 1.
\end{aligned}$$

hence (4.4) is satisfied. Clearly $P \in \Sigma$, so it only remains to show that P is a semimeasure. Since the domains R_i partition ω we have

$$\begin{aligned}
\sum_{x\in\omega} P(x) &= \sum_i \sum_{x\in R_i} P(x) \\
&\leqslant \sum_i \sum_{x\in R_i} P_0(x)2^x \\
&= \sum_i \sum_{x\in R_i} 2^{-x-1} \\
&= \sum_{x\in\omega} 2^{-x-1} = 1.
\end{aligned}$$

□

For some Σ-semimeasures there is a universal Σ-sumtest.

Proposition 4.1.12 ([7]). *Given any computable function $d : \omega \to \omega$, the Σ-semimeasure*

$$P(x) = m(x)2^{-d(x)}$$

satisfies:

- *d is (additively) universal for $S_P \cap \Sigma$.*
- *P is (multiplicatively) universal for the class*

$$\{P' \in \Sigma : d \text{ is } P'\text{-sum-test}\}.$$

Proof. For the first item, suppose that d' is a sum-test for P that is not additively dominated by d, i.e. $d' - d$ is unbounded. Then $P'(x) = m(x)2^{d'(x)-d(x)}$ is a Σ-semimeasure that is not multiplicatively dominated by m, contradicting Theorem 2.2.4. For the second item, suppose that P' is a Σ-semimeasure for which d is a sum-test. Then $Q(x) = P'(x)2^{d(x)}$ is a Σ-semimeasure, hence by Theorem 2.2.4, $P(x)2^{d(x)} = m(x)$ multiplicatively dominates $Q(x)$, and hence $P(x)$ multiplicatively dominates $P'(x)$. □

In subsequent sections, it will be shown that there are no universal Π-sumtests for a universal semimeasure, and a direct product of two universal semimeasures. It is conjectured that this is valid in more generality.

Conjecture 4.1.13. *If a Σ-semimeasure P dominates $Q(x) = 2^{-l(x)-2\log l(x)}$, then P has no universal Π-sumtest.*

4.1.3 Kolmogorov complexity characterizations of Π-sumtests for a Σ-semimeasure

For every Σ-semimeasure P, and computable function f, a Π-sumtests d_f is defined using Kolmogorov complexities. It is shown in Proposition 4.1.17 that any Π-sumtest is dominated by some sumtest d_f for f large enough. This determines how large tests in $S_P^{\downarrow}$ can grow.

Let P be some *fixed* Σ-semimeasure, thus $K(P) \leqslant_+ 0$, and let $S_P^{\downarrow}$, $\dot{S}_P^{\downarrow}$ be short for $S_P \cap \Pi$, $S_P \cap \dot{\Pi}$. For any computable function f, let:

$$d_{f,t}(x) = \min_s\{-\log P_s(x) - K_{f(s)}(x|s) : l(x) \leqslant s \leqslant \max(l(x), t)\}.$$

and let $d_f(x)$ denote the limit for $t = \infty$. By Lemma 4.1.14, d_f defines a Π-sumtest. Let $\exp$ denote the function $\exp n = 2^n$.

Lemma 4.1.14. *For any computable function f: $d_f \in S_P^{\downarrow}$.*

Proof. Since $d_{f,t}(x) = d_{f,l(x)}(x)$ for all $t \leqslant l(x)$, it suffices to show the Lemma for $t \geqslant l(x)$. Remark for such t that

$$-K_{f(t)}(x|t) \geqslant \log P_t(x) + \min_s\{-\log P_s(x) - K_{f(s)}(x|s) : l(x) \leqslant s \leqslant t\}.$$

Therefore

$$\begin{aligned}\sum_{x\in\omega} P(x)\exp d_{f,t}(x) &\leqslant \sum_{x\in\omega}\exp\big(\log P_t(x) \\ &\qquad + \min_s\{-\log P_s(x) - K_{f(s)}(x|s) : l(x) \leqslant s \leqslant t\}\big) \\ &\leqslant \sum_{x\in\omega}\exp -K_{f(t)}(x|t) \\ &\leqslant 1.\end{aligned}$$

□

For any computable function f, let f^* represent a shortest program for f on ϕ. Remark that $l(f^*) = K(f)$. For any computable function f and $w \in 2^{<\omega}$, the notation $w \circ f$ represents the partial computable function computed by $\lambda x : \phi(wf^*|x)$, a function obtained by extending the computation of f with some additional instructions. If $w \circ e$ defines a two argument function $(w \circ e)_t(x)$, then the short notations for the values $(w \circ e)_t(x) = w \circ e_t(x)$ and for fixed t, the notation of the one-argument functions $(w \circ e)_t = w \circ e_t$ is used. It is shown for use in Proposition 4.1.17, that without loss of generality, it can be assumed that for any $e \in \dot{\Pi}$, and for any fixed t, the function e_t "almost" implies a sumtest.

Lemma 4.1.15. *There exists a $w \in 2^{<\omega}$ such that any $e \in \dot{S}_P^{\downarrow}$:*

$$\lim_{t\to\infty} w \circ e_t = \lim_{t\to\infty} e_t,$$

and

$$\sum_{l(x)\leqslant t} P_t(x) \exp w \circ e_t(x) \leqslant 1.$$

Proof. For any t and for increasing $s \geqslant t$, the following sum is evaluated until

$$\sum_{l(x)\leqslant t} P_s(x) \exp e_s(x) \leqslant 1.$$

Since $e \in S_P$, for any t such an $s = s_t$ exists. Let w such that $w \circ e_t = e_{s_t}$. $w \circ e_t$ satisfies the conditions of the lemma. □

Let C be the subset of the computable two-argument functions $f_t(x)$ given by:

$$C = \{f : \forall t\big[\sum_{l(x)\leqslant t} \exp -f_t(x) \leqslant 1\big]\}.$$

Functions $f_t(x)$ in $C \cap \dot{\Sigma}$ can be called compression functions [44], because

$$K(x) =_+ \min_f\{K(f) + f(x) : f \in C \cap \dot{\Sigma}\}.$$

For any computable function g, a similar bound for $K_{g(t)}(x|t)$ is shown using the set $C \cap \dot{\Delta}_2$.

Lemma 4.1.16. *There is a constant w such that for any $f \in C \cap \dot{\Delta}_2$, and for all x, t with $t \geqslant l(x)$:*

$$f_t(x) \geqslant_+ K_{w \circ f(t)}(x|t) - K(f),$$

Proof. Given t, the function values $f_t(y)$ for all $y \leqslant t$ can be evaluated, which contains $f_t(x)$ under the conditions of the lemma. Since $f \in C$, the Shannon-Fano code for all y with $l(y) \leqslant t$ can be constructed from t, which contains a code for x of length $f_t(x) + O(1)$. To decode this Shannon-Fano code, one needs f, t. The necessary computation time for this decoding is given by some function $w \circ f(t)$. □

Proposition 4.1.17. *There is a w such that for any $e \in \dot{S}_P^{\downarrow}$*

$$e \leqslant_+ d_{w \circ e} + K(e).$$

Proof. Without loss of generality it can be assumed that e satisfies the conditions of Lemma 4.1.15. Let

$$h_t(x) = -\log P_t(x) - e_t(x).$$

It follows that $h \in C$. Also, remark that h is limit-computable, therefore $h \in C \cap \dot{\Delta}_2$, and by Lemma 4.1.16 for $t \geqslant l(x)$ one has for some w

$$h_t(x) \geqslant K_{w \circ h(t)}(x|t) - K(e_t) + O(1).$$

Let s be minimal such that $d_{w \circ h, s}(x) = d_{w \circ h}(x)$, then

$$\begin{array}{lll} e(x) & \leqslant & e_s(x) \\ & = & \min\{e_t(x) : l(x) < t \leqslant s\} \\ & \leqslant & \min\{-\log P_t(x) - h_t(x) : l(x) < t \leqslant s\} \\ & \leqslant_+ & \min\{-\log P_t(x) - K_{w \circ e_t(t)}(x|t) \\ & & : l(x) < t \leqslant s\} + K(e) \\ & =_+ & d_{w \circ e, s}(x) + K(e) \\ & =_+ & d_{w \circ e}(x) + K(e). \end{array}$$

□

4.2 Π-sumtests for a universal semimeasure

Since Σ-sumtests for a universal semimeasure are bounded by a constant, the question can be raised how large Π-sumtests for a universal semimeasure can be. In this section it is shown that they can exceed $\log l(x) - O(\log \log l(x))$, but are upper bounded by $\log l(x) + O(\log \log l(x))$. Furthermore, it is shown that they have no universal element. As an intermezzo, length conditional randomness is discussed, and in the last subsection, also an upper bound in terms of minimal set sufficient statistics is given. For the remaining of this chapter, $m = Q_K$ will be chosen.

The notation "$f \leqslant^+ g$" and "for all x: $f(x) \leqslant^+ g(x)$" means that $\exists c \forall x \big[f(x) \leqslant g(x) + c\big]$. In contrast with the "$\leqslant_+$" notation, the implicit c constant may depend

on variables assumed in the context of the statement. More formal, let $QuRx$ denote a series $Q_1u_1Q_2u_2\ldots Q_ku_kRx$ of quantifiers $Q_i, R \in \{\forall, \exists\}$ over the variables u_i, $i \leqslant k$, that are implicitly or explicitly stated in the context of an equation. The expressions "$f \leqslant_+ g$" and "$f \leqslant^+ g$" mean

$$\exists c Q u \forall x \qquad \Big[f_u(x) \leqslant g_u(x) + c\Big]$$
$$Q u \exists c \forall x \qquad \Big[f_u(x) \leqslant g_u(x) + c\Big].$$

In a similar way, $f(x) \longrightarrow^+ g(x)$ means that $K(g(x)|f(x)) \leqslant^+ 0$. Therefore, a similar correspondence between the expressions "$f(x) \longrightarrow_+ g(x)$" and "$f(x) \longrightarrow^+ g(x)$" exists. Constants c implicit in the $O()$ notation are assumed to satisfy $c \leqslant_+ 0$.

4.2.1 Upper bounds for tests in $S^{\downarrow}_m$

It is shown in Proposition 4.2.5 that tests in $S^{\downarrow}_m$ are upper bounded by $\log l(x) + O(\log^{(2)} l(x))$. This result is obtained using Lemma 4.2.1, which shows that if a test in $S^{\downarrow}_m$ is larger than $2k$, for some x, then for any t with $K_t(t) \leqslant k$, then for some computable f: $K_t(x) - K_{f(t)}(x) \geqslant k$. Since there are 2^k such t, one shows that k must be logarithmic. An (f, k)-sequence of times t is defined, which satisfies the required properties for the proof, and which is at the same time also suitable for the proof in Subsection 4.1.1 that shows that the upper bound is tight within an $O(\log\log l(x))$ term.

Lemma 4.2.1. *For $P = m$ and suitable ϕ, there is a w such that for all f, x, t with $t \geqslant l(x)$:*

$$d_f(x) \leqslant_+ K_t(x) - K_{f(t)}(x) + K_{t+1}(t).$$

Proof. For $P = m$ the hierarchy of Π-sumtests d_f is represented by

$$d_{f,t}(x) = \min_s\{K_s(x) - K_{f(s)}(x|s) : l(x) \leqslant s \leqslant \max(l(x), t)\}.$$

For suitable ϕ one has by Proposition 2.4.5, and by $t \longrightarrow K_{t+1}(t)$ that

$$K_{f(t)}(x) \leqslant_+ K_{f(t)}(x, t) \leqslant_+ K_{f(t)+1}(x|t) + K_{t+1}(t).$$

For any $t \geqslant l(x)$ it follows that

$$\begin{array}{rcl} d_f(x) & \leqslant & K_t(x) - K_{f(t)+1}(x|t) \\ & = & K_t(x) - K_{f(t)}(x) + K_{f(t)}(x) - K_{f(t)+1}(x|t) \\ & \leqslant_+ & K_t(x) - K_{f(t)}(x) + K_{t+1}(t). \end{array}$$

□

A sequence of times t will now be fixed with low $K_t(t)$.

Definition 4.2.2. For any $k \in \omega$ and for any computable increasing function f such that for all t, the function $f(t) > t$, the (f,k)-*sequence* $s_0, \ldots, s_e$ is the finite sequence obtained by the subset of

$$z_1 = k2^{k+1},\ z_2 = f(z_1),\ \ldots,\ z_i = f(z_{i-1}),\ \ldots$$

for which there is a t such that $z_i \leqslant t < f(z_i)$ and $K_t(t) < k$.

The following Lemma summarizes some easy observations of (f,k)-sequences, for use in Subsections 4.2.3 and 4.2.4.

Lemma 4.2.3. *Let $s_0, ..., s_e$ be an (f,k)-sequence, then*

(i) *$s_0, \ldots, s_e$ can be enumerated by f, k.*

(ii) *s_i is increasing in i*

(iii) *for all $i < e$: $f(s_i) \leqslant s_{i+1}$,*

(iv) *for all $t \geqslant s_0$ such that $K_t(t) < k$, there is an $i \leqslant e$ such that $s_i \leqslant t \leqslant f(s_i)$,*

(v) *for suitable ϕ and k large enough $s_0 = k2^{k+1}$,*

(vi) *$e < 2^k$.*

Proof. $(i) - (iv)$ follow directly. (v) follows by observing that for suitable ϕ:

$$K_{z_0}(z_0) \leqslant^+ K_k(k) \leqslant^+ \log k + 2\log\log k,$$

(vi) follows since there are maximally $2^k - 1$ programs $p \in 2^{<k}$. □

Lemma 4.2.4. *For suitable ϕ: $k - 2\log k \leqslant^+ \log e$.*

Proof. Assign a free command w on ϕ such that

$$\phi_t(wp|x) = \begin{cases} \max\{\phi_t(p|x), t[p|x]\} & \text{if } \phi_t(p|x)\downarrow \\ \infty & \text{otherwise.} \end{cases}$$

Since the evaluation of the right hand side only requires evaluations of $\phi_t(p|x)$ and since $p < wp$, it follows that this command is well-defined. Remark that for any computable function f the computation time of $w \circ f$ is not larger than $w \circ f$ since $w \circ f = w \circ w \circ f$. It can additionally be assumed on $w \circ f$ that $w \circ f$ is increasing, and $w \circ f(t) > t$ The finite sequence $h_i = (w \circ f)^{(i)}(s_0)$ for $i = 1, \ldots, N$ with $N = 2^{k - 2\log k - c_f}$, and c_f large enough, satisfies for suitable ϕ, and c'_f large enough:

$$\begin{aligned} K_{h_i}(h_i) &\leqslant K_i(i) + c'_f \\ &< N + 2\log N + c_f \\ &\leqslant k. \end{aligned}$$

Also, remark that $h_{i+1} \geqslant f(h_i)$, and thus for any $i \leqslant N$ there is a j such that $h_i \leqslant s_j < h_{i+1}$. This shows that there are at least N different s_j's. □

Proposition 4.2.5. *For all $d \in \dot{S}^{\downarrow}_m$:*

$$d(x) \leqslant^+ \log l(x) - O(\log\log l(x)).$$

Proof. Suppose the proposition is true for a well-defined indexed universal interpreter, then it holds for any indexed universal interpreter, because if $d \in \dot{S}^{\downarrow}_m$, and m' is a second universal Σ-semimeasure, then there is a constant such that $d \in \dot{S}^{\downarrow}_{m'}$. Therefore a suitable ϕ can be assumed anywhere in the proof.

By Proposition 4.1.17 it suffices to show the proposition for each d_f with f satisfying the conditions in Definition 4.2.2.

For any x let $s_0, \ldots, s_e$ be an (f,k)-sequence, with $d_f(x) > k =^+ d_f(x)$, and k small enough such that for any t: $K_t(t) < k$ implies that $K_t(x) - K_{f(t)}(x) \geqslant 1$. Remark that by Lemma 4.2.1 such k exists. Since $f(s_i) \leqslant s_{i+1}$ and using Lemma 4.2.4, it follows that:

$$\begin{aligned} l(x) + O(l(x)) &\geqslant K_0(x) \\ &\geqslant \sum K_{s_i}(x) - K_{f(s_i)}(x) \\ &\geqslant \exp k - 2\log k - c_f \\ &\geqslant \exp d_f(x) - 2\log d_f(x) - c_f, \end{aligned}$$

where c_f is a constant depending on f. Therefore

$$d_f(x) \leqslant l(x) - O(\log l(x)) - c_f.$$

□

4.2.2 Intermezzo: length conditional randomness

Randomness[4] for real numbers can be defined using martingales, which correspond to betting strategies in an iterated game (see also in Chapter 6. In the same way sumtests can also be interpreted as betting strategies for a single game. For a computability class F and a length conditional semimeasure P, a string $x \in 2^n$ is (F,P,c)-*random* if for any $d \in F \cap S_P$:

$$d(x|n) - K(d) \leqslant c.$$

It can be shown that for any constants c, there is a c' such that for computable P and any $x \in 2^n$: (Σ,P,c)-randomness of x implies $K(x|n) \geqslant -\log P(x|n)$, implies (Σ,P,c')-randomness of x. The questions from Subsection 4.1.1 are now connected to the questions on the relation of (Σ,P,c)-randomness and (Π,P,c)-randomness.

[4]This subsection was inspired by comments of an anonymous referee of [5].

Question 4.2.6. *For any c, is there a constant c' such that (Σ, P, c)-randomness implies (Π, P, c)-randomness, for P either in Δ_1 or in Σ ?*

Let m and U be the universal and uniform length conditional, then Proposition 4.2.7 shows that for any c there is a c' such that (Σ, U, c)-randomness implies (Π, m, c')-randomness.

Proposition 4.2.7. *For every $d \in S^{\downarrow}_m$: $K(x|n) \geqslant_+ n$ implies $d(x) \leqslant_+ 0$.*

Proof. For any computable function f, and $x \in 2^n$, the length conditional variant of $d_{f,t}(x)$ for $m(x|n)$ is defined as

$$d_{f,t}(x|n) = \min_s\{K_s(x|n) - K_{f(s)}(x|n, s) : n \leqslant s \leqslant t\}.$$

Remark that the length conditional analogue of Proposition 4.1.17 holds: there is a w such that for any length conditional Π-sumtest e for $m(x|n)$ one has

$$e \leqslant_+ d_{w \circ e} + K(e).$$

Also, remark that the length conditional analogue of Lemma 4.2.1 holds, which implies for $t \geqslant l(x) = n$ that

$$d(x) \leqslant_+ K_t(x|n) - K_{f(t)}(x|n) + K_t(t|n).$$

Choosing $t = n$ implies $K_t(t|n) =_+ 0$ and since

$$K_t(x|n) =_+ K_{f(t)}(x|n) =_+ n,$$

it follows that $d(x) \leqslant_+ K(d)$. □

4.2.3 Logarithmic tests in $S^{\downarrow}_m$

For any function f, the functions $\tilde{d}_f$ are constructed. It is shown that for suitable ϕ, $\tilde{d}_f$ define Π-sumtests for m. Using (f, k)-sequences a gradual compressible x will be constructed. For such x, the function $K_t(x)$ decreases stepwise with increasing t. This x is used in the proof of Proposition 4.2.10 which shows that there are tests in $S^{\downarrow}$ exceeding $\log l(x) - O(\log \log l(x))$.

For any computable function f and some constant $c \in \omega$ large enough, let:

$$\begin{aligned} h(s) &= K_s(s) - 4 \log K_s(s)) - c \\ \tilde{d}_{f,t}(x) &= \min\{K_s(x) - K_{f(s)}(x) + h(s) : l(x) \leqslant s \leqslant \max\{l(x), t\}\}. \end{aligned}$$

The function $\tilde{d}_f$ is obtained by taking the limit $t \to \infty$. The following Lemma shows that $\tilde{d}_f \in \dot{S}^{\downarrow}$, which also shows that the upper bounds in Lemma 4.2.1 are tight within $O(\log \log l(x))$ terms.

Lemma 4.2.8. *For any computable function f, large enough: $\tilde{d}_f \in \dot{S}^{\downarrow}_m$.*

Proof. Let t_k be defined as in Subsection 3.1.2. For suitable ϕ one has

$$K_{t_k}(t_k) \leqslant^+ k + 2\log k.$$

Choosing c in the definition of $\tilde{d}_{f,t}$ large enough implies

$$h(t_k) \leqslant k - 2\log k. \tag{4.6}$$

Let k' be the (m,m)-sophistication, thus:

$$k'(x) = \min\{k : K_{t_k}(x) = K(x)\}.$$

By Corollary 3.1.19: $k'(x) - 2\log k'(x)$ is a sumtest for m. This shows that:

$$\begin{array}{rcl} \tilde{d}_f(x) & \leqslant & K_{t_{k'(x)}}(x) - K_{f(t_{k'(x)})}(x) + h(t_{k'(x)}) \\ & \leqslant & h(t_{k'(x)}) \ \leqslant \ k'(x) - 2\log k'(x). \end{array}$$

This implies that $d_{f,t}$ is a sumtest. □

To have a large $\tilde{d}_f$ one needs a gradual compressible x such that $K_t(x) - K_{f(t)}$, is large, each time $h(t)$ drops to a low value. A nice property of the formalism of indexed interpreters is that for suitable ϕ, the value of $K_t(x)$ is assumed to decrease instantaneously at a predefined t. Lemma 4.2.9 shows that a similar result can also be supposed for a series of times.

A *k-family* of sequences is a set of sequences given by $s_{k,0}, \ldots, s_{k,e_k}$ for each $k \in \infty$. A k-family of sequences is enumerable if there is a partial recursive function f such that for all k and $i \leqslant e_k$ one has $f(i,k) = s_{k,i}$. Let the set 2^n denote the set of all binary strings of length n.

Lemma 4.2.9. *Let $0 < s_{k,0}, \ldots, s_{k,e_k}$ and $l_{k,0}, \ldots, l_{k,e_k} > 0$ be enumerable k-families of increasing and decreasing sequences. For suitable ϕ and for any k there is an $x \in 2^{l_{k,0}}$ such that*

$$K_{s_{k,i}}(x|k,i) \geqslant^+ l_{k,i} \tag{4.7}$$

$$K_{s_{k,i}+1}(x|k,i) \leqslant^+ l_{k,i+1}. \tag{4.8}$$

Proof. The proof below is given for any k. It is assumed that all constants implicit in the notation $\geqslant^+$ and $\leqslant^+$, do not depend on the parameter k.

Let $x^{(k,e)} \in 2^{l_{k,e}}$ such that $K(x^{(k,e)}|k,e) \geqslant l_{k,e}$. For $1 \leqslant i \leqslant e_k$, the string $x^{(k,i-1)} \in 2^{l_{k,i}}$ is inductively defined from $x^{(k,i)}$. Suppose that $x^{(k,i)} \in 2^{l_i}$ is defined, then $x^{(k,i-1)}$ is defined as the lexicographic $x^{(k,i)}$-th string $y \in 2^{l_{i-1}}$,

such that $K_{s_{i-1}}(y|k,i) \geqslant l_{i-1} - 1$. Remark that there are at least $2^{l_{i-1}} - 1$ such strings y, and therefore such a string $x^{(k,i)}$ exists.

By definition:

$$K_{s_{k,i}}(x^{(k,i)}|k,i) \geqslant l_{k,i}. \tag{4.9}$$

It will now be shown that for suitable ϕ and for any $i < e$:

$$K_{s_{k,i}+1}(x^{(k,i)}|k,i) \leqslant l_{k,i+1}. \tag{4.10}$$

Indeed, all strings $y \in 2^{l_{k,i}}$, incompressible in time $s_{k,i}$, can be enumerated, and the $x^{(i+1)}$-th can be chosen. For suitable ϕ this shows the inequality.

Let $x = x^{(0)}$. For any $i < e$, equations (4.8) and (4.7) are now shown, using equations (4.10) and (4.9).

- The reasoning to show equation (4.10) can be iterated for $j \leqslant i$ to show that $x^{(i+1)}$ computes $x^{(0)} = x$ in time $s_{k,i} + 1$ for suitable ϕ. This shows inequality (4.8).

- From $l_{k,0}, s_{k,0}$, all elements $y \in 2^{l_{k,0}}$ can be enumerated with $K_{s_{k,0}}(y|i) \geqslant l_{k,0}$. The lexicographic index of $x = x^{(0)}$ in this enumeration is $x^{(1)}$. In the same way, one computes $x^{(k,i)}$ from $x^{(k,i-1)}$, using $k, i \longrightarrow s_{k,i}$. When iterated one computes $x^{(k,i)}$ from x, i using only evaluations of $\phi_{s_{k,i-1}}(.|.)$.

  Let p be the witness of $K_{s_{k,i}}(x)$. The computation of x from p in time $s_{k,i}$, and the computation of $x^{(k,i)}$ from $x^{(0)} = x$ can be combined in a well-defined command for $\phi_{s_{k,i}}(wps^*l^*|k,i) = x^{(k,i)}$, for s^*, l^* the partial recursive functions witnessing the enumerability of the k-families $s_{.,.}$ and $l_{.,.}$. This implies for suitable ϕ that

$$K_{s_{k,i}}(x|k,i) \geqslant^+ K_{s_{k,i}}(x^{(k,i)}|k,i) \geqslant l_{k,i}.$$

□

Proposition 4.2.10. *There exists a $d \in S_m^{\downarrow}$ such that for any n there is an $x \in 2^n$ such that:*

$$d(x) \geqslant^+ \log n - O(\log^{(2)} n).$$

Proof. The proposition will be shown for any n such that there is a k with $n = k2^{k+1}$. Suppose that the proposition is shown for any such n, then the proposition follows for any other n by padding the constructed x with zeros and adapting d correspondingly. This is explained in more detail: for any n, let k be maximal such that $n \geqslant k2^{k+1} = s_{k,0}$, and let x, d be the corresponding $x \in 2^{s_{k,0}}$ and $d \in S_m^{\downarrow}$. Let $x' \in 2^n$ be obtained from $x \in 2^{s_{k,0}}$ by appending zeros, and let d' be defined from d, such that $d'(x') = d(x)$ for all such x, by ignoring the last $n - s_{k,0}$

bits of x'. Since $\log s_0 =^+ \log n$, this proves the proposition. Therefore it suffices to prove the proposition for $n = k2^{k+1}$, for any k.

Let $s_{k,0}, \ldots, s_{k,e}$ be an (f,k)-sequence, let $n = s_{k,0} = k2^{k+1}$, let $l_{k,i} = n - 2ik$, and let x as in Lemma 4.2.9. Remark that by Lemma 4.2.3, $e < 2^k$, therefore one has $K_0(k,i) \leqslant k + 4\log k$. It follows from

$$\begin{aligned} K_{s_{k,i}}(x) &\geqslant^+ K_{s_{k,i}}(x|k,i) \\ &\geqslant^+ l_{k,i} \\ &\geqslant^+ K_{s_{k,i}+1}(x|k,i) + 2k \\ &\geqslant^+ K_{s_{k,i}+1}(x) - K_0(k,i) + 2k \\ &\geqslant^+ K_{s_{k,i}+1}(x) - k + 4\log k \end{aligned}$$

that

$$K_{s_{k,i}-1}(x) - K_{s_{k,i}}(x) \geqslant^+ k - 4\log k.$$

Using Lemma 4.2.3, for any t with $K_t(t)$ there is an $s_{k,i}$ with $s_{k,i} \leqslant t < f(s_{k,i})$. Remind that $f^{(2)}(t) = f(f(t))$. It follows that

$$\begin{aligned} \tilde{d}_{f^{(2)}}(x) &=^+ \min_t\{K_t(x) - K_{f^{(2)}(t)}(x) + h(t)\} \\ &\geqslant \min_t\{K_t(x) - K_{f^{(2)}(t)}(x) : K_t(t) < k\} \cup \{h(t) : K_t(t) \geqslant k\} \\ &\geqslant \min_t\{K_{s_{k,i}}(x) - K_{f(s_{k,i})}(x) : i \leqslant e\} \cup \{k - 4\log k\} \\ &\geqslant k - O(\log k) \\ &\geqslant \log n - O(\log\log n). \end{aligned}$$

□

4.2.4 There is no universal element in $S^{\downarrow}$

Lemma 4.2.11. *Let x, f as in the proof of Proposition 4.2.10, and let g such that for any $f(s) = g(s+1)$ then*

$$d_g(x) \leqslant^+ O(\log k).$$

Proof. The Lemma follows by Lemma 4.2.1 for

$$t = l(x) + 1 = s_{k,0} + 1 = k2^{k+1} + 1$$

by showing that

$$K_0(t) \leqslant O(\log k) \tag{4.11}$$

$$K_t(x) - K_{g(t)}(x) \leqslant O(\log k). \tag{4.12}$$

Since the function $k \to k2^{k+1} + 1$ is primitive recursive, one has for suitable ϕ that

$$K_0(t) \leqslant^+ \log k + 2\log\log k,$$

which shows equation (4.11). Since $g(t) = f(s_0) \leqslant s_1$, one has

$$K_{g(t)}(x|0,k) \geqslant^+ l_1.$$

Therefore for suitable ϕ

$$\begin{aligned} K_t(x) &\leqslant_+ K_{s_{k,0}+1}(x) \\ &\leqslant^+ l_1 + 2\log k \\ &\leqslant^+ K_{g(t)}(x|0,k) + 2\log k \\ &\leqslant^+ K_{g(t)}(x) + 4\log k, \end{aligned}$$

which shows equation (4.12). □

Proposition 4.2.12. *For any $d \in S_m^{\downarrow}$ there is a $d' \in S_m^{\downarrow}$ such that there are infinitely many $x \in 2^{<\omega}$ with:*

$$d'(x) - d(x) \geqslant \log l(x) - O(\log\log l(x)).$$

Proof. By Lemma 4.2.11. □

Corollary 4.2.13. *There is no universal element in $S^{\downarrow}$.*

Proof. By definition. □

4.2.5 An upper bound by the length of a minimal sufficient statistic.

Remind Definition 3.1.21 of a minimal sufficient statistic of a binary string x [14, 44, 71]:

$$l_c^S(x) = \min\{K(S) : x \in S \text{ and } K(x) \geqslant K(S) + \log|S| + c\}.$$

Proposition 4.2.14 shows that functions in $S^{\downarrow}$ can be interpreted as Π-lower bounds for the length of a minimal sufficient statistic within small terms. However, since tests in $S_m^{\downarrow}$ are upper bounded by $\log l(x) + O(\log\log l(x))$, and $l_c^S(x)$ can equal $l(x) - c$ for some constant c, it follows that in general, $l_c^S(x)$ is can not be approximated by some Π-function.

Proposition 4.2.14. *For all c and $d \in S^{\downarrow}$: $d(x) \leqslant^+ l_c^S(x) + 4\log l_c^S(x)$.*

Proof. Let S be the set that realizes the definition of $l_c^S(x)$. x can be computed using S in a time s: enumerate all elements of S and let i be the index of x in this enumeration. Therefore $S^* \longrightarrow s$. Generating x in this way requires a program of length $K(S) + \log|S| \leqslant_+ K(x) + c$ and a computation time s that is bounded by a computable function of $t[S^*]$ and the elements of S. For this s we have

$K_s(x) \leqslant_+ K(x) + c$. Let c' be the constant such that $K_s(x) \leqslant K(x) + c'$, then it follows that $s \geqslant t_{k_{c'}}$. Therefore:

$$S^*, k_{c'} \longrightarrow s, k_{c'} \longrightarrow t_{k_{c'}},$$

and thus

$$l_c^S(x) = K(S) \leqslant_+ K(t_{k_{c'}}) - 2 \log l_c^S(x).$$

For $t = t_{k_{c'}}$, and for f large enough one has by Proposition 4.1.17 and Lemma 4.2.1:

$$\begin{aligned} d(x) &\leqslant^+ d_f(x) \leqslant^+ K_t(x) - K^{f(t)}(x|t) + K_t(t) \\ &\leqslant^+ c + K_t(t) \leqslant^+ l_c^S(x) + 2 \log l_c^S(x). \end{aligned}$$

□

4.3 Independence tests

The set of independent distributions over $\omega \times \omega$ is given by all $P(x, y)$ such that there exist Q, R with $P(x, y) = Q(x)R(y)$. In the following section it will be shown that these semimeasures have a universal element, given by $m(x)m(y)$, for any universal semimeasure m.

Many results about sum-tests from the previous sections hold, mutates mutandis, for the case of independence tests, with the same proofs. In particular, in the case of $P = Q = m$, Corollary 4.1.5 now states that there are no unbounded computable and Σ-independence tests. There exist unbounded Π tests, and we will show that there is no Π-universal test by Proposition 4.2.12. As a corollary to the proof it follows that for all Σ-semimeasures P, Q, a Π-independence test for (P, Q) exist, with $d(x, y) \geqslant l(x) - O(\log l(x))$ for infinitely many binary strings x, y with length $l(x) = l(y)$, and for each Π-independence test d for (m, m), there is a test d' such that $d'(x, y) - d(x, y)$ exceeds $l(x) - O(\log l(x))$ infinitely often. Since $P = Q = m$ throughout this section, "independence test" will abbreviate "independence test for m and m".

We start with an informal argument why there is no Π-universal independence test. Consider the set

$$D = \big\{(x, y) : l(x) = l(y) \wedge x, y \text{ random and dependent}\big\}.$$

D is a natural example of a d.c.e. set, that is, a set that is the difference of two c.e. sets, in this case the set of pairs (x, y) with x and y dependent minus the set of pairs where one of x and y is not random. Now suppose that d is a Π-independence test. It follows directly from the definition of independence tests, that the set of pairs x, y where $d(x, y)$ is large, is small in measure. Thus d provides us with

an effective method for detecting dependencies in such pairs. Now suppose that for all $(x, y) \in D$, $d(x, y)$ would be large. Then we would have that x and y are dependent if and only if $d(x, y)$ is large. Since the latter is a Π-event, we obtain that $D \in \Pi$, a contradiction. This means that there are $(x, y) \in D$ such that $d(x, y)$ is small, that is, x and y are dependent but d does not see this. Since D is a set of small measure, we could construct a new d' with d' higher on such pairs (thus showing that d is not universal). To recognize such pairs, we have to recognize more dependencies than d does by allowing for more computation time. Some pairs (x, y) may fall through at a later time when it turns out that one of x and y is not random, but if we allow for enough computation time we will also find pairs in D that were not recognized by d, and hence we can show that d is not universal. The proof below is more informative, since it shows that the functions d^i of the specific form defined there form a strict hierarchy of independence tests, and that every independence test is dominated by some d^i.

Proposition 4.3.1. *There is no universal* Π*-independence test.*

Proof. By Proposition 4.1.17, for any Π-independence test e, there is an f such that $e \leqslant^+ d_f$. It can be additionally assumed that f exceeds its computation time. The propositions follows now by constructing for every n and computable f, strings $x, y \in 2^n$, such that

$$\begin{aligned} d_{f+5}(x, y) &\geqslant n - O(\log n) \\ d_f(x, y) &\leqslant O(\log n). \end{aligned}$$

This will follow from the length conditional variants of the sumtests

$$d_{f+5}(x, y|n) \geqslant n \tag{4.13}$$

$$d_f(x, y|n) \leqslant^+ 0. \tag{4.14}$$

Remark that changing the choice of m changes the set of sumtests to some sets of sumtests that maximally differ by an additive constant. Therefore the choice $m_s(x) = 2^{-K_s(x)}$ is used ($m = Q_K$). The sumtests d_f can now be written as

$$\begin{aligned} d_{f,t}(x, y) = \min\{&K_s(x) + K_s(y) - K_{f(s)}(x, y|s) \\ &: \max(l(x), l(y)) \leqslant s \leqslant \max(l(x), l(y), t)\}. \end{aligned}$$

Equation (4.13) and (4.14) follow by substituting strings x, y as in Lemma 4.3.3, and choosing $t = l(x) = n$ in the above equation. □

Lemma 4.3.2. *For suitable* ϕ*, for all* n*, and for computable* f*, exceeding its computation time, there exist strings* $a, x \in 2^n$ *such that:*

- $K_{f(n)+4}(a|n) \leqslant^+ 0$

- $K(x|n) \geqslant^{+} n$
- $K_{f(n)+2}(a|x) \geqslant^{+} n$.

Proof. Let c be a large enough constant. Let a be the lexicographic first string of length n that cannot be produced from n by a program of length less than n in time less than $f(n)+3$. There is always such a string a. One can compute a from n by a description of f. Since f exceeds its own computation time, for suitable ϕ this can happen such that the first condition of the Lemma is satisfied.

There is at least one $x \in 2^n$ with $K(x|a) \geqslant n$. Note that $K(x|n) \geqslant^{+} K(x|a) \geqslant n$, and by this the second condition is satisfied.

By definition of a and x one has

$$2n \leqslant^{+} K_{f(n)+3}(a|n) + K(x|a).$$

For suitable ϕ and for $s = f(n)+1$ in Lemma 2.4.5 it follows that

$$2n \leqslant^{+} K_{f(n)+2}(x|n) + K_{f(n)+2}(a|x).$$

For suitable ϕ it holds that $K_{f(n)+2}(x|n) \leqslant^{+} n$, hence

$$2n \leqslant^{+} n + K_{f(n)+2}(a|x).$$

By this, the last condition is satisfied. □

Lemma 4.3.3. *For suitable ϕ, for all n, and for computable f exceeding its computation time, there are $x, y \in 2^n$ satisfying:*

$$\begin{aligned} K(x|n) &\geqslant^{+} n \\ K(y|n) &\geqslant^{+} n \\ K_{f(n)}(x,y|n) &\geqslant^{+} 2n \\ K_{f(n)+5}(x,y|n) &\leqslant^{+} n. \end{aligned}$$

Proof. For any n large enough, pick x and a as in Lemma 4.3.2. Therefore, the first inequality of the Lemma is satisfied. Let $y = \mathrm{XOR}(x,a)$, where XOR is the bitwise exclusive-or operator. The remaining inequalities are now shown.

- Note that $\mathrm{XOR}(y,a) = \mathrm{XOR}(\mathrm{XOR}(x,a),a) = x$. This provides a program for x given a and y. It follows that $K(x|n) \leqslant^{+} K(y|n) + K(a|y)$ and hence:

$$\begin{aligned} K(y|n) &\geqslant^{+} K(x|n) - K(a|y) \\ &\geqslant^{+} K(x|n) - K_{f(n)+4}(a|n) \\ &\geqslant^{+} K(x|n) \\ &\geqslant^{+} n. \end{aligned} \tag{4.15}$$

  This shows the second inequality of the Lemma.

- Since $\mathrm{XOR}(y,x)=a$, it follows that any program computing y from x, also computes a from x. For suitable ϕ one has

$$K_{f(n)+1}(y|x) \geqslant^{+} K_{f(n)+2}(a|x) \geqslant^{+} n. \tag{4.16}$$

  Furthermore we have $K_{f(n)}(x|n) \geqslant^{+} n$. Hence, for suitable ϕ and $s = f(n)$ in Proposition 2.4.5 one has

$$K_{f(n)}(x,y|n) \geqslant^{+} K_{f(n)+1}(x|n) + K_{f(n)+1}(y|x) \geqslant^{+} 2n. \tag{4.17}$$

  This shows the third inequality of the Lemma.

- Since $\mathrm{XOR}(x,a)=y$, for suitable ϕ

$$K_{f(n)+5}(y|x) \leqslant^{+} K_{f(n)+4}(a|x) \leqslant^{+} 0.$$

  Therefore

$$K_{f(n)+5}(x,y) \leqslant^{+} K_{f(n)+5}(x) + K_{f(n)+5}(y|x) \leqslant^{+} n. \tag{4.18}$$

  This shows the forth inequality of the Lemma.

□

The proof of Proposition 4.3.1 is actually stronger, and implies some corollaries. First remark that for a length conditional variant of the independence tests, the proposition shows that for any $d \in S^{\downarrow}$, there is a d' such that for any n and some $x,y \in 2^n$ one has that $d(x,y|n) \leqslant^{+} 0$, while for $d'(x,y|n) \geqslant^{+} n$. It is d considers x,y maximally independent, while d' considers x,y maximally dependent.

Corollary 4.3.4. *Algorithmic mutual information*

$$I(x;y|n) = K(x|n) + K(y|n) - K(x,y|n)$$

is within additive $O(\log n)$ terms an independence test that dominates all length conditional Π-independence tests.

Proof. Since d_f does so. □

Corollary 4.3.5. *There exists a constant c, such that for all Σ-semimeasures P,Q, there exist a Π-independence test d for P,Q such that $d(x,y|n) \geqslant^{+} n$ for some $x,y \in 2^n$.*

Proof. For some i large enough, there are infinitely many x, y with $l(x)=l(y)$ and

$$d_f(x,y|n) \geqslant^{+} n.$$

By universality of m, we have that $P(x) \leqslant 2^{c_P} m(x)$ and $Q(x) \leqslant 2^{c_Q} m(x)$, for some constants c_P, c_Q. Remark that $d(x) = d^i(x) - c_P - c_Q$ defines an independence test for P and Q. □

From the proof it also follows that

Corollary 4.3.6. *There is a constant c, such that for all Π-independence tests d, there is a Π-independence test d' with*

$$d'(x, y|n) - d(x, y|n) \geqslant^{+} n,$$

for infinitely many x, y with $l(x) = l(y) = n$.

Proof. Note that for $i = K(d) + O(1)$ we have

$$d^i(x, y) - d(x, y) \geqslant n - c \log n - c_i.$$

Hence for all n with $\log n \geqslant c_i$ we have

$$d^i(x, y) - d(x, y) \geqslant n - (c + 1) \log n.$$

□

5

Statistically explanatory models and influence tests

Abstract. The relation between structural modeling and universal semimeasures defined with objective probabilities is discussed. This leads to the definition of the hypotheses of causality and influence for discrete time series. Several ideal tests are introduced, and it is argued that when Halting information is transmitted, in some cases, instantaneous cause and consequence can be inferred where this is not possible classically.

The approach is contrasted with Bayesian definitions of influence, where it is left open whether all Bayesian causal associations of universal semimeasures are equal within a constant.

Finally the approach is also contrasted with existing engineering procedures for influence and alternative theoretical definitions of causation.

An extensive literature exists on the definition of 'influence' both in statistics and in philosophy [13, 41, 45, 53, 66]. However, most of this work defines influence only when a fixed probability distribution for some observables is already available. General purpose tests are considered here, that define influence without reference to a semimeasure. Such definitions are useful to interpret algorithms such as [20, 34, 36, 52, 54, 57, 61, 74]. Until now, there is no theory available that considers both statistical interpretations and computability aspects.

Causality is often related to structural equations [53]. Traditionally, computable functions are used to study these generalized structural equations. The set of semimeasures corresponding to these structural equations do not lead to sets

of semimeasures with a universal element. To solve this problem, structural equations with partial computable functions will be considered.

The logarithm of the proposed ideal statistical tests, define an algorithmic variant of the Shannon information transfer. In section 5.3 both quantities are related, and therefore an alternate interpretation of the algorithms in [52,54,57,74,74] can be given. Also, Granger causality can be interpreted in this framework. Finally, the proposed test for influence is contrasted to Shannon information transfer of the minimal sufficient statistic, and to graphical representations of minimal sufficient statistics as in [42]. It is shown that Σ-algorithmic information transfer determines plausible causal relations, where these relations can not be determined from probabilistic minimal sufficient statistics.

Length conditional semimeasures are used, which allows to reduce technical details. Furthermore they can be justified by remarking that in many experimental setups, the amount of generated data, is fixed before the experiment starts.

5.1 Explanatory models and causal semimeasures

This section derives several influence tests from generalized structural equations both for pairs of discrete variables, and for pairs of discrete time series. It is shown that when Halting information is present in two discrete observations x, y, the obtained universal elements from the structural equation hypotheses can imply slightly different likelihoods if x is assumed to cause y or y is assumed to cause x. When x, y represent discrete time series, the difference in likelihood can become significant depending on whether x is assumed to instantaneously cause y, or y is assumed to instantaneously cause x.

5.1.1 Statistical explanatory model

Many hypotheses can be defined in terms of (generalized) structural equations, this is especially true for causal models [53]. To relate a structural equation to semimeasures the objective uncertainty formalism described in Subsection 1.1.3, no "preference" may exist for the value of the hidden or uncontrolled variable in the structural equation. By relating structural equations to semimeasures in this way, the corresponding group of semimeasures can have a universal Σ-element. For the hypotheses described in future subsections, the additional assumption that there might be no preference on the different values of the hidden or uncontrolled observable in the structural equations, does not influence the universal Σ-semimeasure corresponding to the structural equations.

Cantor space $2^\omega = [0, 1]$ with tree topology is assumed, it is, the open sets are determined by any $r \in 2^{<\omega}$ as

$$[r] = \{\alpha \in 2^{<\omega} : r \sqsubset \alpha\},$$

with $r \sqsubset \alpha$ meaning that r is a prefix of α. Also, the measure $\mu([r]) = 2^{-l(r)}$ will be assumed on this topology.

Let $X \in \omega$ denote a discrete observable, a *statistical explanatory model* for X, is given by some unobservable, or uncontrolled variable $R \in [0,1]$ with a probabilistic description given by a semimeasure P_R over the unit interval $[0,1]$, and some function f such that $X = f(R)$. For some observation x of the observable X, if $x = f(r)$, then f, r is a probabilistic explanation of the observed data x, where r represents the hidden or uncontrolled variables of the context where the value x of X is observed. The a priori probability of occurrence of x is given by:

$$P_{f,R}(x) = \int dr\{r : x = f(r)\},$$

where Lebesgue integration over r, with respect to the measure P_R is performed, and if f is assumed to be integrable.

For many contexts, it can be assumed that f is partial recursive and P_R is a Σ-semimeasure. According to Lemma 5.1.1, the set of semimeasures obtained by assuming such general R to be distributed over such a P_R, and assuming R uniformly distributed over $[0,1]$, is equivalent.

Lemma 5.1.1. *If the variable R is distributed according to a Σ-semimeasure P_R, and if f is partial computable, then there is a partial computable f' and a uniform distributed variable R' on $[0,1]$, such that for all x:*

$$P_{f,R}(x) = P_{f',R'}(x).$$

Proof. First the function α is inductively defined. For any x, let $\alpha(x,0) = 0$ and let

$$\alpha(x,t) \quad = \quad \alpha(2^n, t-1) + \sum\{P_t(z) - P_{t-1}(z) : z \leqslant x\}.$$

Remark that for every r' such that

$$0 \leqslant r' \leqslant \sum\{P_{f,r}(x) : x \in 2^n\},$$

there is a unique t such that $r' \in [\alpha(2^n, t-1), \alpha(2^n, t)]$. Therefore, each such r' defines a unique x, such that

$$r' \in [\alpha(x-1,t), \alpha(x,t)].$$

If $l(r')$ is long enough, then also

$$[r'] \in [\alpha(x-1,t), \alpha(x,t)].$$

is satisfied, it is

$$r' \in [\alpha(x-1,t), \alpha(x,t) - 2^{l(r')}]. \tag{5.1}$$

Let $f'(r')$ be the function that is defined to be x if there is an x such that (5.1) is satisfied, and undefined otherwise. Remark that f' is partial computable and satisfies the conditions of the Lemma. □

From now on, the variable R will be assumed to have the uniform distribution over $[0,1]$, and P_f is short for $P_{f,R}$. According to Proposition 5.1.2, the set of explanatory models is equivalent with the set of Σ-semimeasures.

Proposition 5.1.2. *For every partial computable f, the semimeasure P_f is a Σ-semimeasure. For every Σ-semimeasure P, there is a partial computable function f such that $P = P_f$.*

Proof. The first claim follows by definition. Therefore, the second claim remains to be proved. Let $\alpha(x,t)$ be as in the proof of Lemma 5.1.1. The second claim follows by choosing $f(r) = x$ if there is an x such that for some t

$$\alpha(x,t) \leqslant r \leqslant \alpha(x,t) + 2^{-l(r)},$$

and $f(r) = \infty$ (undefined) otherwise. Remark that f is partial computable, and satisfies the conditions of the Lemma. □

Let t_i as in section 3.1. Remind that t_i increases faster than any computable function on i, the probability for a prefix-free Turing machine that a program halts after time t_i is bounded by $o(2^{-i})$. It can be shown that only for a small measure of hidden and uncontrolled variables R, there are x'es for which the exploratory model needs more computation time than t_i.

5.1.2 Causal explanations for a pair of observables

Different types of explanatory models are defined, and the corresponding universal elements are compared.

- An explanatory model for two discrete observables X, Y is given by a partial computable function f_{XY} and a variable R, uniformly distributed over $[0,1]$, such that:
$$(X,Y) = f_{XY}(R).$$
- An explanatory model for two *independent* discrete observables X, Y is given by two partial computable functions f_X, f_Y and two variables R_X, R_Y, independently and uniformly distributed over $[0,1]$, such that:
$$\begin{aligned} X &= f_X(R_X) \\ Y &= f_Y(R_Y). \end{aligned}$$

- An explanatory model for two discrete variables X, Y such that X *causes* Y is given by two partial computable functions $f_X, f_{Y|X}$ and two variables R_X, R_Y, independently and uniformly distributed over $[0, 1]$, such that

$$X = f_X(R_X) \tag{5.2}$$
$$Y = f_{Y|X}(X, R_Y). \tag{5.3}$$

Proposition 5.1.3. *The semimeasures corresponding to the explanatory models for X, Y, respectively, independent X, Y, and X causing Y, have universal elements, which are given by $m(x, y)$, respectively, $m(x), m(y)$, and $m(x|y)m(y)$.*

Proof. This follows in a similar way as Proposition 5.1.2, and by Proposition 2.2.6. □

Let x^* be a program of length $K(x)$ that computes x, then by the Coding Theorem and additivity of K, (Corollary 2.4.6), it follows for any bivariate universal semimeasure that

$$m(x, y) =^* m(y|x^*)m(x). \tag{5.4}$$

The test for the hypothesis that x is independent from y if x is a probabilistic cause of y is given by:

$$\frac{m(x)m(y|x)}{m(x)m(y)} = \frac{m(y|x)}{m(y)}.$$

The test for the hypothesis that x is independent from y if x, y are generated in the most general way, is given by:

$$\frac{m(x, y)}{m(x)m(y)} = \frac{m(y|x^*)}{m(y)}, \tag{5.5}$$

by equation 5.4. Remark that to approximate this test, a shortest model for x might be needed. By Proposition 5.1.4 these tests can differ.

Proposition 5.1.4. *For every n and all $x, y \in 2^n$*

$$\frac{m(y|x^*)}{m(y|x)} \leqslant^* n.$$

For every n, there are $x, y \in 2^n$ such that

$$\frac{m(y|x^*)}{m(y|x)} \geqslant^* \frac{n}{\log n}$$
$$\frac{m(x|y)m(y)}{m(y|x)m(x)} \geqslant^* \frac{n}{\log n}$$

Proof. The first claim of the proposition follows by first applying the conditional variant of the Coding Theorem 2.4.3 to:

$$\begin{aligned} \log m(y|x^*) &=^+ K(y|x^*) \\ \log m(y|x) &=^+ K(y|x). \end{aligned}$$

Remark that by Lemma 2.4.7, one has $K(x), x \longrightarrow_+ x^*$, and by [44, page 242] it follows that

$$K(x^*|x) =_+ K(K(x)|x) \leqslant^+ \log n.$$

Remark that $K(y|x) \leqslant_+ K(y|x^*) + K(x^*|x)$. Combining the above equations shows the claim.

The second claim of Proposition 5.1.4 is now shown. Remark that $K(x)$ can be computed from x^*, and that

$$K(x) =_+ K(K(x), x) =_+ K(K(x)) + K(x|K(x)^*).$$

Let $y = K(x)$. By applying the conditional Coding Theorem, it only needs to be shown that there exists for every n some $x \in 2^n$ such that

$$K(K(x)|x) \geqslant_+ K(K(x)|x^*) + \log n - \log\log n$$

Remark that $K(K(x)|x^*) =_+ 0$. By [44, Theorem 3.8.1], it follows that for every n there is an $x \in 2^n$ with $K(K(x)|x) \geqslant^+ \log n - \log\log n$. □

Corollary 5.1.5. *There are hypotheses S, T such that for every n, there are $x, y \in 2^n$ such that*

$$\frac{m^{(S\times T)^\uparrow}(x,y)}{m^{S^\uparrow \times T^\uparrow}(x,y)} \geqslant^* \frac{n}{\log n}$$

Proof. Let S be the hypothesis that x is generated by a partial computable function of a hidden variable r_x, and let T be the hypothesis that y is generated from x by any function of a hidden variable r_y and x. The universal element of $S^\uparrow$ is given by $m(x)$, the universal element of $T^\uparrow$ is given by $m(y|x)$. By Proposition 2.2.6, the universal element of $S^\uparrow \times T^\uparrow$ is given by $m(x)m(y|x)$.

It will now be shown that the universal element of $(S\times T)^\uparrow$ is given by $m(x,y)$. First remark that the semimeasure $m(x,y)/m(x)$ corresponds to some generalized structural equations where y is generated from x and a hidden variable r_y, by some function $f(x,r)$ that is not partial computable. Since $m(x,y)$ is also computable $m(x,y) \in (S \times T)^\uparrow$, but since $m(x,y)$ is also universal to the most general enumerable set of semimeasures, it must be universal to $(S \times T)^\uparrow$.

Finally it needs to be shown that for every n there are $x, y \in 2^n$ such that:

$$\frac{m(x,y)}{m(x)m(y|x)}.$$

This follows from Proposition 5.1.4 and equation 5.4. □

Remark that if the partial computable functions f_* in the definitions of the different exploratory models where chosen computable, the corresponding sets would in general not have a universal element. Any computable semimeasures $P(x|y), P(y)$, also generate semimeasures $P(y|x), P(x)$ that satisfy $P(x|y)P(y) = P(y|x)P(x)$. This means that it does not matter for the likelihood of x, y to describe first x and then y and vice versa. This contrasts with the likelihood obtained from enumerable universal semimeasure by Proposition 5.1.4. Therefore, when $K(K(x)|x)$ is large, and y contains much information about $K(x)$, the hypothesis that x caused y can be considered more plausible than vice versa, and the likelihood can differ by a factor n.

Before a last type of hypothesis is introduced, the definition of total arguments of partial computable functions is given.

Definition 5.1.6. A partial computable function $f(x, y)$ on $2^{<\omega} \times U$ for some set U has a *total argument* x, iff for any $x \in 2^{<\omega}$ and $y \in U$ such that $f(x, y)$ is defined, also any $f(w, y)$ is defined with $w \in 2^{l(x)}$.

If a function $f(x, y)$ has a total argument x, it is denoted as $f(\underline{x}, y)$.

The hypothesis of y being *totally caused* by x is given by partial computable functions $f_X, f_{Y|X}$, such that $f_{Y|X}(\underline{X}, R_X)$ has a total argument X, and by two variables R_X, R_Y independently and uniformly distributed over $[0, 1]$, satisfying equations (5.2) and (5.3). To relate such explanatory models to semimeasures, total conditional semimeasures are introduced.

Definition 5.1.7. A *total conditional semimeasures* is a semimeasure P such that for all $n \in \omega$, there is a constant $P(\varepsilon|\underline{n})$ such that for all $x \in 2^n$:

$$P(\varepsilon|x) = P(\varepsilon|\underline{n}).$$

By Proposition 2.2.6, the total conditional semimeasures have a universal Σ-semimeasure. Defining causality with this hypothesis can lead to fundamentally different results by the following Lemma.

Lemma 5.1.8. *For all n, there are $x, y \in 2^n$ such that:*

$$\log \frac{m(y|x)}{m(y|\underline{x})} \geqslant n - O(\log n).$$

Proof. The proof uses total conditional prefix-free complexity $K(x|\underline{y})$, which is explained in detail in Section 6.1. The result follows from the standard Coding Theorem, the total Coding Theorem, Proposition 6.1.7, and from the existence of $x, y \in 2^n$ for all n, such that

$$K(x|y) - K(x|\underline{y}) \geqslant^+ n - 2\log n,$$

Proposition 6.1.2. □

This difference is due to Halting information present in y, Proposition 6.1.3. It can be interpreted as follows: if the computation of $f_{Y|X}$ requires a time t_i (see higher) for some large i, then x, r must contain a large amount of Halting information. Remark that t_i contains about i bits of Halting information [6]. For a general partial computable function $f_{Y|X}$, this Halting information can be obtained from both arguments of the function r and x. If $f_{Y|X}$ is total in its first argument, and $f_{Y|X}(x, r)$ is defined, then a program can be made that generates t_i from $f_{Y|X}$ and r. Therefore if the computation of y is so involved that it needs a time t_i, then i bits of Halting information are present in r, and such probability decreases with 2^{-i}. This is not the case for the partial computable $f_{Y|X}$.

5.1.3 Causal and influence-free explanations for two time series

Let $X, Y \in \omega^{<\omega}$ be observables representing time series. The hypothesis that X is an *instantaneous cause* of Y, is defined as the existence of partial computable functions f_X, f_Y, and variables R_X, R_Y uniformly and independently distributed over $[0,1]^n$ such that for all $i \leqslant n$:

$$\begin{aligned} X_i &= f_X(X^{i-1}, Y^{i-1}, R_X^i, n) \\ Y_i &= f_Y(X^i, Y^{i-1}, R_Y^i, n). \end{aligned}$$

See figure 5.1, right, black and red.

The hypothesis that X, Y are *strict causal*, is defined as the existence of partial computable functions f_X, f_Y, and variables R_X, R_Y uniformly and independently distributed over $[0,1]^n$ such that for all $i \leqslant n$:

$$\begin{aligned} X_i &= f_X(X^{i-1}, Y^{i-1}, R_X^i, n) \\ Y_i &= f_X(X^{i-1}, Y^{i-1}, R_Y^i, n). \end{aligned}$$

See figure 5.2, black. Remark that by symmetry, if X is a strict cause of Y, then Y is a strict cause of X.

The hypothesis that X is *influence-free* of Y, is defined as the existence of partial computable functions f_X, f_Y, and variables R_X, R_Y uniformly and independently distributed over $[0,1]^n$ such that for all $i \leqslant n$:

$$\begin{aligned} X_i &= f_X(X^{i-1}, R_X^i, n) \\ Y_i &= f_X(X^i, Y^{i-1}, R_Y^i, n). \end{aligned}$$

See figure 5.1, right, black and red.

The most general structure is obtained if hidden variables are shared. Therefore the hypothesis that X, Y *can have hidden variables* is given by the partial computable functions f_X, f_Y and the variable R uniformly distributed over $[0,1]^n$

such that for all $i \leqslant n$:

$$\begin{aligned} X_i &= f_X(R^i, n) \\ Y_i &= f_Y(R^i, n). \end{aligned}$$

This model is both equivalent with the models from figure 5.1, left, and 5.2, right, black and red.

Remark that also total versions for these hypotheses can be defined.

5.1.4 Causal semimeasures and ratio tests

The hypothesis described in the previous subsection, correspond to sets of Σ-semimeasures which are investigated in this subsection. These semimeasures will be called causal and online semimeasures, corresponding to the total and the non-total of the hypothesis of instantaneous causality from the previous section. The study of the total versions is motivated philosophically, by remarking that the total semimeasures defined below make more sense for length conditional semimeasures, whereas the online semimeasures makes more sense for subjective interpretations of probability as in [12]. First the total versions are studied.

For $x \in 2^n$ and $i \leqslant n$, let

$$P(x^i|n) = \sum\{P(x^i v|n) : v \in 2^{n-i}\},$$

and similar for $P(x^i, y^j|n)$ and $P(x^i|y)$. For $k \leqslant i \leqslant n$ and $l \leqslant j \leqslant n$, let

$$P(x^i, y^j|x^k, y^l, n) = \frac{P(x^i, y^j|n)}{P(x^k, y^l|n)}.$$

To save notation, for a semimeasure P and for $x, y \in 2^n$, let $P(x^i) = P(x^i|n)$, and more general let $P(x^i, y^j|x^k, y^l, n)$.

Definition 5.1.9. Let $x, y \in 2^n$.

- The causal semimeasure and the instantaneous causal semimeasure, *associated* with a conditional semimeasure $P(x|y)$ is given by:

$$\begin{aligned} P(x\lceil y\uparrow) &= \prod\{P(x_i|x^{i-1}, y^{i-1}) : i \leqslant n\} \\ P(x\lceil y\uparrow^+) &= \prod\{P(x_i|x^{i-1}, y^i) : i \leqslant n\}. \end{aligned}$$

- A conditional semimeasure $P(y|x)$ is *causal* respectively *instantaneous causal*, iff for all $i \leqslant n$ respectively

$$\begin{aligned} P(y\lceil x\uparrow) &= P(y|x) \\ P(y\lceil x\uparrow^+) &= P(y|x). \end{aligned}$$

  Notation: if a semimeasure is causal, or instantaneous causal, it is denoted by $P(x|\underline{y}\uparrow)$, and $P(x|\underline{y}\uparrow^+)$.

- x is *influence-free* of y according to a semimeasure $P(x, y)$, iff:

$$P(x\lceil y\uparrow) = P(x), \tag{5.6}$$

when defined.

Proposition 5.1.10. *For any* ***m****easure $P(x, y)$, the following statements are equivalent:*

(i) $P(y|x)$ *is instantaneous causal.*

(ii) $\forall i \leqslant n \forall x, y \in 2^n \left[P(y^i|x) = P(y^i|x^i)\right]$ *where defined.*

(iii) $\forall i \leqslant n \forall x, y \in 2^n \left[P(x|x^i, y) = P(x|x^i)\right]$ *where defined.*

(iv) $\forall i \leqslant n \forall x, y \in 2^n \left[P(x_{i+1}|x^i, y) = P(x_{i+1}|x^i)\right]$ *where defined.*

(v) x *is influence-free of* y *according to* $P(x, y)$.

Proof. $(i) \rightarrow (ii)$**:**
Let

$$P(y^i|x^i\uparrow^+) = \prod\{P(y_j|y^{j-1}, x^j) : j \leqslant i\}.$$

First it is shown that

$$P(y^i|x) = P(y^i|x^i\uparrow^+). \tag{5.7}$$

Suppose that for some y^i, x: $P(y^i|x) > P(y^i|x^i\uparrow^+)$, then for every $j = i + 1, ..., n$, choose y_j such that $P(y_{j+1}|x, y^j) \geqslant P(y_{j+1}|x^{j+1}, y^j)$. Remark that this is always possible. This shows that:

$$\begin{aligned} P(y|x) &= P(y^i|x)\prod\{P(y_j|x, y^{j-1}) : j = i + 1...n\} \\ &> P(y^i|x^i\uparrow)\prod\{P(y_j|x^j, y^{j-1})\} \\ &= P(y|x\uparrow^+). \end{aligned}$$

which contradicts (i). (ii) follows by

$$\begin{aligned} P(y^i|x^i) &= \sum\{P(y^i|z)P(z|x^i) : z^i = x^i\} \\ &= \sum\{P(y^i|z^i\uparrow^+)P(z|x^i) : z^i = x^i\} \\ &= P(y^i|x^i\uparrow^+). \end{aligned}$$

$(ii) \rightarrow (iii)$**:** By Bayes theorem.
$(iii) \rightarrow (iv)$**:** By summing over all $x^{i+1}v$ with $v \in 2^{n-i-1}$.
$(iv) \rightarrow (v)$**:** By definition.
$(v) \rightarrow (i)$**:** By remarking that

$$P(x\lceil y\uparrow)P(y\lceil x\uparrow^+) = P(x, y) = P(x)P(y|x).$$

□

Remark that the set of causal, and instantaneous causal semimeasures is testable and convex. Therefore, they have universal elements $m(x|\underline{y}\uparrow), m(x|\underline{y}\uparrow^+)$. The hypotheses defined in the previous subsection correspond to sets of semimeasures which have a universal Σ-element.

Proposition 5.1.11. *The universal element of the hypothesis that:*

1. *X is a instantaneous total cause of Y, is given by $m(x|\underline{y}\uparrow)m(y|\underline{x}\uparrow^+)$.*
2. *X, Y are strict total causal, is given by $m(x|\underline{y}\uparrow)m(y|\underline{x}\uparrow)$.*
3. *X is total influence-free of Y, is given by $m(x)m(y|\underline{x}\uparrow)$.*
4. *X, Y have hidden common variables, is given by $m(x, y)$.*

Proof. The corresponding sets of semimeasures are products of convex Σ-sets of Σ-semimeasures. The result follows by Proposition 2.2.6. □

The universal elements define ideal hypotheses tests. Some of them can be simplified within a constant factor, using $m(x, y) =^* m(x)m(y|x^*)$.

- Suppose that X is an instantaneous cause of Y,
  are X, Y strict causal according to data x, y?
  Figure 5.2, left.

$$\frac{m(x|\underline{y}\uparrow)m(y|\underline{x}\uparrow^+)}{m(x|\underline{y}\uparrow)m(y|\underline{x}\uparrow)} = \frac{m(y|\underline{x}\uparrow^+)}{m(y|\underline{x}\uparrow)} \tag{5.8}$$

- Suppose that X, Y are strict causal,
  is Y influence-free of X according to data x, y?
  Figure 5.1, right.

$$\frac{m(x|\underline{y}\uparrow^+)m(y|\underline{x}\uparrow)}{m(x|\underline{y}\uparrow^+)m(y)} = \frac{m(y|\underline{x}\uparrow)}{m(y)} \tag{5.9}$$

- Suppose X, Y can have hidden variables,
  is X an instantaneous cause of Y according to data x, y ?

$$\frac{m(x, y)}{m(x|\underline{y}\uparrow^+)m(y|\underline{x}\uparrow)} \tag{5.10}$$

- Suppose X, Y can have hidden variables,
  are X, Y strict causal ?
  Figure 5.2, right.

$$\frac{m(x, y)}{m(x|\underline{y}\uparrow)m(y|\underline{x}\uparrow)} \tag{5.11}$$

- Suppose X, Y can have hidden variables,
  is Y influence-free of X ?

$$\frac{m(x|y^*)m(y)}{m(x|\underline{y}\uparrow^+)m(y)} = \frac{m(x|y^*)}{m(x|\underline{y}\uparrow^+)} \tag{5.12}$$

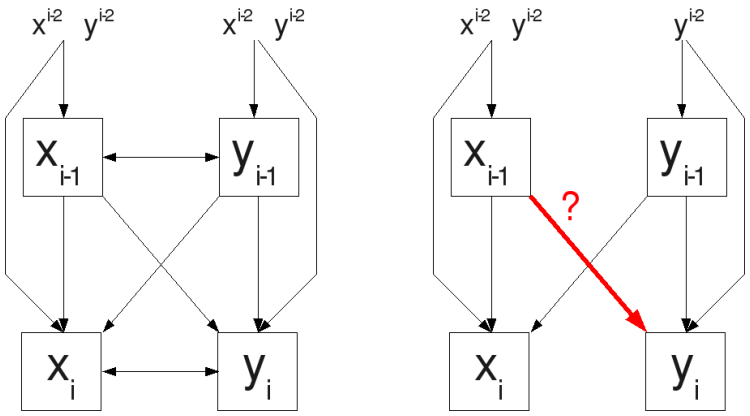


Figure 5.1: Left: general system. Right: suppose that X, Y are strict causal, is Y influence-free of X ?

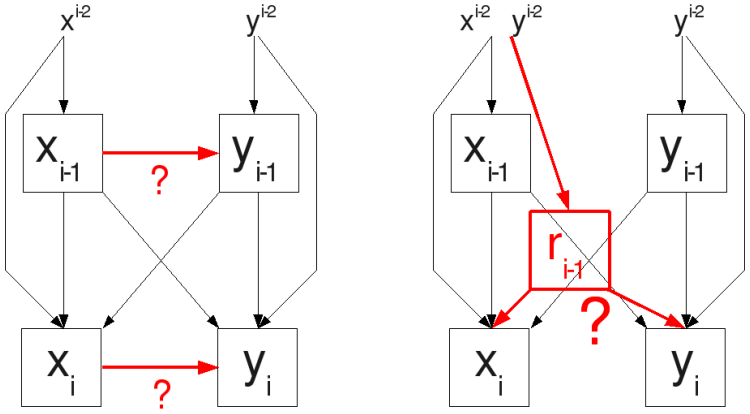


Figure 5.2: Left: Suppose that X is an instantaneous cause of Y, are x, Y strict causal ? Right: Suppose X, Y can have hidden variables, are X, Y strict causal ?

The test that x is independent of y, given that x, y are generated in the most general way, in equation (5.5), can now be written as a product of some tests above. Remind that the test values actually correspond to significances of statistical tests, therefore such products have a nice interpretation, as products of significances.

For example as the product of the tests of equations (5.11), (5.9), and (5.9) applied to x, or as the decomposition. (5.10), (5.8), (5.9), and (5.9) applied to x, or as the decomposition.

$$\begin{aligned} \frac{m(x,y)}{m(x)m(y)} &= \frac{m(x,y)}{m(x|\underline{y}\uparrow)m(y|\underline{x}\uparrow)}\frac{m(y|\underline{x}\uparrow)}{m(y)}\frac{m(x|\underline{y}\uparrow)}{m(x)} \\ &= \frac{m(x,y)}{m(x|\underline{y}\uparrow^+)m(y|\underline{x}\uparrow)}\frac{m(y|\underline{x}\uparrow^+)}{m(y|\underline{x}\uparrow)}\frac{m(y|\underline{x}\uparrow)}{m(y)}\frac{m(x|\underline{y}\uparrow)}{m(x)} \end{aligned} \tag{5.13}$$

The length conditional variant of online semimeasures [12] is now defined.

Definition 5.1.12. A function $P : 2^{<\omega} \times 2^{<\omega} \to [0,1] : (x,y) \to P(x|y,n)$ defines an *(length conditional) online semimeasure* iff for all n, for all $x, y \in 2^n$, for all $i < n$, and for all $b \in \{0,1\}$ one has:

$$P(x^i|y^{i-1}) \leqslant P(x^i 0|y^i b) + P(x^i 1|y^i b),$$

and

$$P(0|y^0) + P(1|y^0) \leqslant 1.$$

Remark that the set of online semimeasures differs from the set of causal semimeasures, since for causal semimeasures, the first defining inequality is always an equality. Remark that the set of online semimeasures are convex and testable, and therefore have a universal element, which is denoted as $m(x|y\uparrow)$. Remark that also instantaneous online semimeasures can be defined. For online semimeasure Proposition 5.1.11 can now be given for online semimeasures in an identical way.

Proposition 5.1.13. *The universal element of the hypothesis that*

1. *X is a instantaneous cause of Y, is given by $m(x|y\uparrow)m(y|x\uparrow^+)$.*
2. *X, Y are strict causal, is given by $m(x|y\uparrow)m(y|x\uparrow)$.*
3. *X is influence-free of Y, is given by $m(x)m(y|x\uparrow)$.*
4. *X, Y have hidden common variables, is given by $m(x,y)$.*

Similar hypotheses tests can now be defined as in the total case.

5.1.5 Σ-information transfer and instantaneous information transfer

Equation (5.13) allows a nice information theoretic interpretation. Let the Σ-information transfers be defined as follows:

$$\begin{aligned} I(x;y) &= \log \frac{m(x,y)}{m(x)m(y)} \\ \Sigma IT(x \leftarrow y) &= \log \frac{m(x|y\uparrow)}{m(x)} \\ \Sigma IT(x\uparrow;\underline{y}\uparrow) &= \log \frac{m(x,y)}{m(x|y\uparrow)m(y|x\uparrow)}. \end{aligned}$$

Equation (5.13) becomes now:

$$I(x;y) = \Sigma IT(x \leftarrow y) + \Sigma IT(y \leftarrow x) + \Sigma IT(x\uparrow;y\uparrow).$$

Suppose that x, y have no instantaneous connections, then the mutual information of x, y, can be considered as the sum of information flowing from the past of x to y, from the past of y to x, and information obtained by x, y through a hidden source.

However, a decomposition of mutual information as the sum of information flowing from the past and the present of x to y and from the past of y to x is not possible. Also, in this setting, there can be a different instantaneous information flow if it is assumed that information is instantaneously flowing from x to y, or from y to x. Both claims follow from Proposition 5.1.14.

Proposition 5.1.14. *For every n there exist $x, y \in 2^n$ such that:*

$$\Sigma IT(x\uparrow;y\uparrow) - \log \frac{m(y|x\uparrow^+)}{m(y|x\uparrow)} \geqslant o(n).$$

For there exist $x, y \in \omega^n$ such that:

$$\log \frac{m(x|y\uparrow^+)}{m(x|y\uparrow)} - \log \frac{m(y|x\uparrow^+)}{m(y|x\uparrow)} \geqslant o(\sum\{\log x_i + \log y_i : i \leqslant n\}).$$

Proof. Follows directly from Lemma 5.1.15. □

Lemma 5.1.15. *There is a constant c such that for any $n \geqslant \frac{1}{c}$, there exist $x, y \in 2^n$ such that*

$$\log \frac{m(x,y)}{m(x|y\uparrow)m(y|x\uparrow^+)} \geqslant cn.$$

Proof. This follows from the online coding theorem, Proposition 6.2.4, and from the non-additivity of online complexities, Proposition 6.3.1. □

Remark that for the total variants, exactly the same conclusions hold.

5.2 Associated causal semimeasures

In the previous section, causal Σ-semimeasures where derived as corresponding to structural equations with partial computable functions. In this section causal semimeasures are investigated that are in a Bayesian way associated to Σ-conditional semimeasures and Σ-bivariate semimeasures.

Definition 5.2.1. A semimeasure $P(x|y)$ is *associated causal* respectively *associated instantaneous causal* if it there is a conditional Σ-semimeasure $Q(x|y)$ such that $P(x|y) = Q(x\lceil y\uparrow)$ respectively $P(x|y) = Q(x\lceil y\uparrow^+)$.

Remark that an Σ-causal semimeasure $P(x|y)$ is associated causal, since it equals its own association. Also, remark that the set of associated causal semimeasures is not convex. Since with any bivariate semimeasure $P(x, y)$, a conditional semimeasure is associated, one can associate a causal and instantaneous causal semimeasures also with $P(x, y)$.

Lemma 5.2.2.

$$P(x, y) = P(\varepsilon, \varepsilon)P(x\lceil y\uparrow)P(y\lceil x\uparrow^+). \tag{5.14}$$

Proof. Remark that for the causal semimeasures $P(x\lceil y\uparrow)$ and $P(y\lceil x\uparrow^+)$ associated with $P(x, y)$ one has

$$\begin{aligned} P(x\lceil y\uparrow) &= \frac{P(x^1, y^0)}{P(x^0, y^0)} \cdots \frac{P(x^n, y^{n-1})}{P(x^{n-1}, y^{n-1})} \\ P(x\lceil y\uparrow^+) &= \frac{P(x^1, y^1)}{P(x^0, y^1)} \cdots \frac{P(x^n, y^n)}{P(x^{n-1}, y^n)}. \end{aligned}$$

□

5.2.1 Non existence of universal elements

In contrast with the causal Σ-semimeasures, the associated causal semimeasures have no universal element.

Proposition 5.2.3. *For any conditional Σ-semimeasure $P(x|y) > 0$, there exists a conditional Σ-semimeasure $Q(x|y)$ and $x, y \in 2^n$ such that*

$$\log \frac{Q(x\lceil y\uparrow)}{P(x\lceil y\uparrow)} > o(n). \tag{5.15}$$

Proof of Proposition 5.2.3 first part: definition of Algorithm 1.
Let $N = 2n$ and let the set $2^n \times 2^n$ be associated with 2^N by mapping x, y to $z = x_1y_1x_2y_2...x_ny_n$. With abuse of notation P, denote the restriction of P on 2^N with P. For any such restricted semimeasure P and $v \in 2^i, i \leqslant N$, let $P(v...)$

denote the restriction of P on the strings vu for all $u \in 2^{N-i}$. For $b \in \{0,1\}$, let $\overline{b} = 1 - b$. The strings of 2^N can be considered as branches in a tree. For $z \in 2^N$, z is a local minimal branch, iff it satisfies for all $i \leqslant n$:

$$P(z^i) \leqslant P(z^{i-1}\overline{z_i}).$$

For a local minimal branch z, the nodes $z^{2i+1}\overline{z^{2i+2}}$ for $i \leqslant n-1$ are called *load* nodes. Algorithm 1 generates for every restriction P on 2^n of a computable semimeasure a computable semimeasure Q on 2^n such that all leafs w have half weight, it is $Q(w) = P(w)/2$, except for leafs descending from load nodes which have $Q(w) = P(w)$. This implies that the weights of the uneven local minimal nodes are proportionally more heavy than the weights of the even local minimal nodes, which shows the result of Lemma 5.2.4.

Lemma 5.2.4. *If $P(x|y\uparrow)$ is a computable causal semimeasure associated with a computable semimeasure $P(x,y) > 0$, then $Q = grow(P)$, with algorithm grow defined in Algorithm 1, is computable and satisfies:*

$$\log \frac{Q(x|y\uparrow)}{P(x|y\uparrow)} \quad > \quad o(n).$$

Proof. Algorithm 1 constructs Q from P such that:

$$Q(w) = \begin{cases} P(w) & \text{if } w \text{ is a load leaf,} \\ \frac{1}{2}P(w) & \text{otherwise.} \end{cases}$$

For $i < N$ and z the local minimal leaf,

$$P(z^{i+1}) \leqslant \frac{1}{2}P(z^i),$$

and for $i < n$,

$$\begin{aligned} Q(z^{2i}) - \frac{1}{2}P(z^{2i}) &= Q(z^{2i+1}) - \frac{1}{2}P(z^{2i+1}) \\ &= Q(z^{2i+2}) - \frac{1}{2}P(z^{2i+2}) \\ &\quad + Q(z^{2i+1}\overline{z_{2i+2}}) - \frac{1}{2}P(z^{2i+1}\overline{z_{2i+2}}) \\ &\geqslant \frac{1}{2}P(z^{2i+1}\overline{z_{2i+2}}) \geqslant \frac{1}{4}P(z^{2i+1}). \end{aligned}$$

This shows that:

$$\begin{aligned}
\frac{Q(z^{2i+1})}{Q(z^{2i})} &= \frac{\frac{1}{2}P(z^{2i+1}) + Q(z^{2i+1}) - \frac{1}{2}P(z^{2i+1})}{\frac{1}{2}P(z^{2i}) + Q(z^{2i}) - \frac{1}{2}P(z^{2i})} \\
&\geqslant \frac{\frac{1}{2}P(z^{2i+1}) + \frac{1}{4}P(z^{2i+1})}{\frac{1}{2}P(z^{2i}) + \frac{1}{4}P(z^{2i+1})} \\
&\geqslant \frac{P(z^{2i+1})}{P(z^{2i})} \frac{1+\frac{1}{2}}{1+\frac{1}{2}P(z^{2i+1})/P(z^{2i})} \\
&\geqslant \frac{6}{5}\frac{P(z^{2i+1})}{P(z^{2i})}.
\end{aligned}$$

$$\begin{aligned}
-\log Q(x|y\uparrow) &= \sum_{i\leqslant n} -\log \frac{Q(z^{2i+1})}{Q(z^{2i})} \\
&= \sum_{i\leqslant n} -\log \frac{P(z^{2i+1})}{P(z^{2i})} - \log\frac{6}{5} \\
&\geqslant -\log P(x|y\uparrow) - n\log\frac{6}{5}.
\end{aligned}$$

Remark that Algorithm 1 constructs a computable Q from a computable P. □

Data: P
Result: Q
begin
  $z \longleftarrow$ local minimal branch in 2^N
  $Q \longleftarrow \frac{1}{2}P$
  for i *from* 0 *to* $n-1$ **do**
    $Q(z^{2i+1}\overline{z^{i+2}}...) \longleftarrow P(z^{i+1}\overline{z^{i+2}}...)$ (load node)
end

algorithm 1: grow

Proof of Proposition 5.2.3 second (last) part.
The causal semimeasure associated with a bivariate semimeasure $P(x,y)$, is the causal semimeasure associated with the conditional semimeasure $P(x|y)$. For every conditional Σ-semimeasure $P(x|y) > 0$, there is an Σ-semimeasure $Q(x,y) > 0$ such that $Q(x|y) = P(x|y)$. Therefore, to show Proposition 5.2.3, it suffices to show the proposition for causal semimeasures associated with bivariate semimeasures.

For computable semimeasures, the proposition is proved by Lemma 5.2.4. By Lemma 5.2.5, it follows that when some Q_t satisfies equation 5.15 then also Q_{t+s}

Data: P_t
Result: Q_t
begin
$\nu = \min\{P_0(w) : w \in 2^N\}$
$Q_0(w) \longleftarrow \text{grow} 2^{-1/\nu} P_0(w)$
$S \longleftarrow P_0(\varepsilon)$
$s \longleftarrow 0$
for t *from* 0 *to* ∞ **do**
if $P_t(\varepsilon) - S > \nu$ **then** stage s: new Q_t is grown
$S \longleftarrow P_t(\varepsilon)$
$s \longleftarrow s + 1$
$Q_t \longleftarrow \text{grow}(2^{-1/\nu+s} P_t)$
else
$Q_t \longleftarrow \frac{P_t}{P_{t-1}} Q_{t-1}$
end

algorithm 2: grow_semimeasure

satisfies this equation, if Q_{t+s} has not grown to much. Algorithm 2 maintains a Q_t from P_t by growing a Q_0 from P_0, and for $t > 0$ performing a proportional update of Q_t. Each time P_t has grown substantially, it grows a new Q_t. This is possible by taking $Q_0(w)$ very small for any $w \in 2^N$.

Let $\nu = \min\{P_0(w) : w \in 2^N\}$. Remark that the enumeration P_t can be chosen such that $\nu > 0$, since $P(x, y) > 0$. Algorithm 2 uses Algorithm 1, to define an Σ-semimeasure Q_t from the Σ-semimeasure P_t. It is now shown that $Q_{t+1} \geqslant Q_t$: suppose that $t, t+1$ are in the same stage s, then this is easily observed, if at time $t+1$ a new stage $s+1$ is reached, a new Q_{t+1} is grown from P_{t+1} multiplied with a factor $2^{1/\nu+s+1}$, which is doubled relative to the previous stage. Therefore if w was a non-load leaf at time t, and becomes a load leaf at time $t+1$, one still has $Q_{t+1}(w) \geqslant Q_t(w)$. By Lemma 5.2.4 it follows for every t that initiates a new stage, that:

$$\frac{Q_t(x|y\uparrow)}{P_t(x|y\uparrow)} \geqslant o(n).$$

By Lemma 5.2.5, this equation also hold for the t subsequent to the t's initiating a new stage. Therefore, the equation holds for any t. □

Lemma 5.2.5. *Suppose that for some $\nu > 0$, and some semimeasures $P, Q \in 2^N$, one has*

$$Q(\varepsilon) \leqslant P(\varepsilon) + \nu,$$

and for all $w \in 2^N$,

$$Q(w) \geqslant P(w) \geqslant \nu,$$

then

$$\frac{1}{2} \leqslant \frac{Q(x\lceil y\uparrow)}{P(x\lceil y\uparrow)} \leqslant 2.$$

Proof. Since any branch of depth j has 2^{N-j} leafs, one has that $P(w^j) \geqslant \nu 2^{N-j}$.

$$\begin{aligned}
Q(x\lceil y\uparrow) &= \prod\{\frac{Q(w^{2i+1})}{Q(w^{2i})} : i < n\} \\
&\geqslant \prod\{\frac{P(w^{2i+1})}{P(w^{2i}) + \nu} : i < n\} \\
&\geqslant \prod\{\frac{P(w^{2i+1})}{P(w^{2i}) + P(w^{2i})2^{-N+2i}} : i < n\} \\
&\geqslant P(x|y\uparrow)\prod\{\frac{1}{1+2^{-2i}} : i < n\} \\
&\geqslant \frac{1}{2}P(x|y\uparrow) \\
Q(x\lceil y\uparrow) &= \prod\{\frac{Q(w^{2i+1})}{Q(w^{2i})} : i < n\} \\
&\leqslant \prod\{\frac{P(w^{2i+1}) + \nu}{P(w^{2i})} : i < n\} \\
&\leqslant \prod\{\frac{P(w^{2i+1}) + P(w^{2i+1})2^{-N+2i+1}}{P(w^{2i})} : i < n\} \\
&\leqslant P(x|y\uparrow)\prod\{1 + 2^{-2i-1} : i < n\} \\
&\leqslant 2P(x|y\uparrow)
\end{aligned}$$

□

5.2.2 Causal semimeasures associated with a universal semimeasure

Let $m(x\lceil y)$ be the causal semimeasure associated with $m(x, y)$ and let $\tilde{m}(x\lceil y)$ be the causal semimeasure associated with $m(x|y)$.

Conjecture 5.2.6. *There is a constant $c > 0$ such that for all $n \geqslant \frac{1}{c}$, there are $x, y \in 2^n$ with*

$$\log \frac{m(x\lceil y\uparrow)}{m(x|y\uparrow)} \geqslant cn,$$

and there are $x, y \in 2^n$ with

$$\log \frac{\tilde{m}(x\lceil y\uparrow)}{m(x|y\uparrow)} \geqslant cn.$$

Conjecture 5.2.7. *$m(x\lceil y\uparrow)$ and $\tilde{m}(x\lceil y\uparrow)$ are not in Σ.*

Question 5.2.8. [1] *Let S be the set of causal semimeasures associated with an universal Σ-semimeasure. How much can two elements of S differ. Has S a universal element ?*

The relations betweens the sets of associated and causal Σ-semimeasures are represented in figure 5.3.

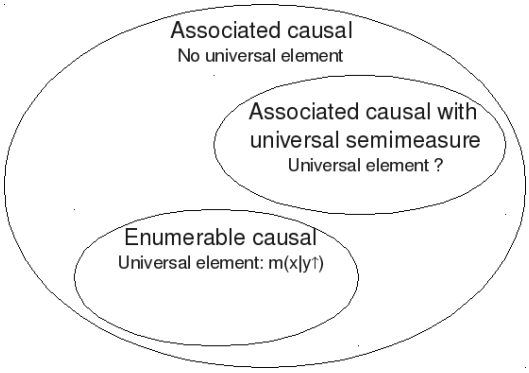


Figure 5.3: Conjectured relations between sets of causal semimeasures and existence of universal elements.

5.2.3 Associated information transfer and instantaneous common information

Associated information transfer and instantaneous common information are given by:

$$\begin{aligned} AIT(x \leftarrow y) &= -\log \frac{m(x\lceil x\uparrow)}{m(x)} \\ AIT(x\uparrow; y\uparrow) &= -\log \frac{m(x,y)}{m(x|y\uparrow)m(y|x\uparrow)}. \end{aligned}$$

Remark that:

$$I(x;y) = AIT(x \leftarrow y) + AIT(y \leftarrow x) + AIT(x\uparrow; y\uparrow).$$

[1] A serious unsuccessful attempt has been made by the author, the problem seems to be very hard.

Associated simultaneous information transfer has also another representation.

$$AIT(x\uparrow; y\uparrow) = \log \frac{m(x\lceil y\uparrow^{+})}{m(x\lceil y\uparrow)} = \log \frac{m(y\lceil x\uparrow^{+})}{m(y\lceil x\uparrow)}.$$

This means that in contrast with the Σ-instantaneous common information, the associated instantaneous common information can be interpreted as the sum of an instantaneous information flow from x to y, a flow from y to x, and a simultaneous flow from a hidden source to x and y.

5.3 Shannon information transfer and minimal sufficient statistics

5.3.1 Granger causality and Shannon information transfer

Statistical tests used in engineering literature can often be structured as follows: first a model is fitted on the data and then influence is derived from:

- Some parameters in the model, as for example by the use of directed transfer functions [36] and partial directed coherences [56].

- The complexity or some mean magnitude of the noise of the data relative to the model, as for example with the use of Granger Causality [17, 20, 27, 34, 36], and Shannon information transfer [52, 54, 57, 74].

By an on-line Coding Theorem, Theorem 6.2.4 and Proposition 6.2.3, the ideal statistical tests based on Σ-information transfer, can be informally assumed to derive influence from the sum of the complexity of the model, and the complexity of the noise. It is not clear whether such algorithms perform better [8].

Let $E(X^{+}|X^{-})$ denote some average error of a prediction strategy of observations of the observable X given its past observations. Let $E(X^{+}|X^{-}, Y^{-})$ be similar where the prediction strategy also uses the past of Y. In its most general form [20, 27], Y is said to Granger causal X iff

$$E(X^{+}|X^{-}) - E(X^{+}|X^{-}Y^{-}))$$

is large. The most common choice for $E(.|.)$ to define Granger causality is the mean squared error.

Another choice for $E(.|.)$ is Shannon entropy. The following expressions provide definitions for Shannon mutual information, information transfer and instan-

taneous mutual information:

$$\begin{aligned} SI_P(X;Y) &= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\ SIT_P(X \leftarrow Y) &= \sum_{x,y} P(x,y) \log \frac{P(x|y\uparrow)}{P(x)} \\ SIT_P(X\uparrow; Y\uparrow) &= \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x|y\uparrow)P(y|x\uparrow)}. \end{aligned}$$

Remark that

$$SI_P(x;y) = SIT_P(x \leftarrow y) + SIT_P(y \leftarrow x) + SIT_P(x\uparrow; y\uparrow).$$

A general procedure of deriving influence for the procedures in [52, 54, 57, 74] is given by fitting some models $P(X_t|X_{t-k\ldots t-1})$, $P(X_t|X_{t-k\ldots t-1}, Y_{t-k\ldots t-1})$, to the corresponding data segments and similar for Y_t, and finally computing the statistic $SIT_P(X \leftarrow Y) - SIT_P(X \leftarrow Y)$. A confidence for the sign of the statistic can be obtained by running the procedure on some randomized permutation of the sequences x and y.

The continuous entropy of a Normal distribution is given by $\sqrt{2\pi e}\sigma$ [58]. This implies that when the error of the observed data relative to some model is assumed to be Normal distributed, the Shannon entropy is estimated by the root mean squared error, in correspondence with common definitions of Granger causality.

When it is assumed that P is a good model for the data, in a frequentist interpretation, this means that for repetitive observation of the data, the data is distributed according to P, then ideal on-line data compression is with overwhelming probability performed by Shannon-Fano code [44]. The expected difference of the code-length of the on-line Shannon-Fano code and the unconditioned code is given by the Shannon information transfer. The expected code-length for optimal on-line encoding is given by AIT transfer within small terms. Therefore, mean Shannon information transfer and mean Σ-information transfer are equal within some constant. A formal version of this statement is given by Proposition 5.3.1.

Proposition 5.3.1.

$$\sum \sum_{x,y\in\omega} P(x,y) AIT(x \leftarrow y) \quad =^+ \quad SI_P(X \leftarrow Y) \pm K(P)$$

Proof. The proof is similar as in [26, Lemma II.4]. □

Since the "probability of the undefined" relative to a computable function is small [64], the following is result is conjectured.

Conjecture 5.3.2.

$$\sum_{x,y\in\omega} P(x,y) AIT(x \leftarrow y) =^+ \sum_{x,y\in\omega} P(x,y) \Sigma IT(x \leftarrow y) \pm K(P)$$

5.3.2 Minimal sufficient statistics and ideal Shannon information transfer

Algorithms for extracting P from x, y as in the previous subsection, are often designed to let P model as much as possible properties that appear frequently within the time series, while at the same time keeping the descriptional complexity of P low.

To idealize this procedure, it has been conjectured [42] that in the case of multivariate models, the constructed P should be chosen as a probabilistic minimal sufficient statistic of the data x, y. Two ways of assigning causal relationships from such a P exists, either by computing SIT_P or by extracting a graphical schema from P. In [42] it is argued informally that for the multivariate case, a minimal Bayesian network is a minimal sufficient statistic. At [42, Lemma 4], it is claimed that if a two-part code satisfying some syntactical form results in an incompressible string, the first part is the probabilistic minimal sufficient statistic. However, it is argued here that in many cases, a plausible graphical causal representation cannot be contained in a probabilistic minimal sufficient statistic of the data, while ΣIT does reveal the plausible causal structure.

Proposition 5.3.3. *For some c, and for any n, there are strings x, y and x', y' for which the same P is a minimal c-sufficient statistic such that*

$$\begin{array}{lllll} \Sigma IT(x \leftarrow y) & =^+ & 0 & \Sigma IT(x' \leftarrow y') & =^+ \quad n - \log n \\ \Sigma IT(y \leftarrow x) & =^+ & n - O(\log n) & \Sigma IT(y' \leftarrow x') & =^+ \quad 0. \end{array}$$

Proof. For some Solovay function f, let n be such that $K(n) =_+ f(n)$. Let x, y as constructed in the proof of Proposition 4.3.1, however, a is not chosen as the lexicographic first string incompressible in time $f(n)$, but rather, incompressible in time t_k. Equations 4.17, remains, and equation 4.18 becomes now

$$K_{t_k+3}(x, y|n) \leqslant^+ \leqslant^+ n + \log n.$$

Let P_{k+1} as defined in equation (3.1). Remark that $K(P_{k+1}|n) =_+ k$, and therefore it defines a minimal length conditional c-sufficient statistic. However, by the Solovay property, it also defines a minimal c-sufficient statistic. Let $x = x' = x''$, and let $y' = 0y^{n-2}$, and let $y'' = y_1 \dots y_{n-1}0$. Remark that both (x, y),(x', y'), and (x'', y'') share the same minimal c-sufficient statistic. □

The argument can be extended to show the analogue of Proposition 5.3.3 for the multivariate case with a complex incompressible causal structure, that contains no Halting information.

Conclusion

Using generalized structural equations, different hypotheses of influence and causality can be defined. These hypotheses define sets of Σ-semimeasures that have a universal element, and therefore they define different statistical tests. The detailed derivation of the statistical tests shows that there can be substantial differences in the corresponding confidences depending on the presumed directions of instantaneous information flows.

Associated causal semimeasures define a larger set of causal semimeasures that are not Σ-functions, nor have a universal element. For the set of semimeasures associated with universal semimeasures, it is not clear whether a universal element exists, and consequently it is not clear whether they define some natural independence tests. However, these tests can define ideal influence tests without assumptions on instantaneous information transfer. Different conjectured relations are summarized in figure 5.3.

Finally the ideal methods of information transfer are contrasted with practical methods from literature. Also, the method is contrasted with the use of minimal sufficient statistics and it is shown that Σ-information transfer can describe plausible causal relations where minimal sufficient statistics can not do.

6
Online Kolmogorov complexity

Abstract. Coding results provide formal motivations to approximate the hypotheses tests defined in Chapter 5 by data compression heuristics. Here the corresponding online and total online length conditional Kolmogorov complexity are defined, and coding results are shown.

Subsequently, it is investigated whether an additivity result is possible, similar to the case of conditional Kolmogorov complexities:

$$K(x, y) = K(x) + K(y|x^*).$$

Using either online and total online variants, and using either conditioning on the witness or not, four possible decomposition candidates are given. It is shown that the non conditional versions differ by an $o(l(x))$ constant. Relations of this result with Muchnik's paradox on online randomness is discussed. It is shown that the total conditional decomposition define an exact decomposition within an additional logarithmic term of the m-sophistication of the strings. The accuracy of the decomposition for the conditional non-total version is left open.

The online complexities define several variants of instantaneous information transfer. The question rises whether these instantaneous information transfer is symmetric. It is shown which definitions are symmetric and which not. However, for many definitions this question is left open.

6.1 Total conditional complexity

Two variants of online complexity will be investigated: online complexity and total online complexity. To understand the difference between these two complexities,

total conditional complexities are studied. Total conditional complexities were also introduced in [49, 59]. Length conditional complexities will be used, rather then monotone complexities, since there are some involved issues, making a nice coding theorem difficult [15, 24].

Definition 6.1.1. For any n, and for $x, y \in 2^n$, the *total conditional complexity* of a string x given y is:

$$K(x|\underline{y}) = \min\left\{l(p) : \forall z \in 2^n[\phi(p|z)\downarrow] \wedge \phi(p|y)\downarrow = x\right\}.$$

6.1.1 Conditional and total conditional complexity

Proposition 6.1.2 shows that total conditional complexity differs substantially from conditional complexity.

Proposition 6.1.2. *For all n there are $x, y \in 2^n$ such that:*

$$\begin{aligned} K(x|\underline{y}) &=_+ n \\ K(x|y) &\leqslant_+ 0. \end{aligned}$$

Proof. Let y be the program with longest computation time t of maximal length n. Remark that $K(y|n) \geqslant_+ n$. Append zeros to y such that $l(y) = n$. Let $x \in 2^n$ be the lexicographic first string that can not be computed from y within time t. Remark that $K(x|y) \leqslant_+ 0$. Let f be a computable function defined on 2^n with $f(y) = x$. The computation time of a minimal program p that evaluates $f(z)$ for all $z \in 2^n$ has computation time larger than t, and therefore $K(f) =_+ l(p) \geqslant_+ K(t) \geqslant_+ n$. □

Proposition 6.1.3 shows that the difference between total conditional and conditional complexity is due to halting information in y. The Kolmogorov complexity relative to the Halting problem K' is defined as the Kolmogorov complexity relative to a universal prefix-free machine with Halting oracle. The Mutual information with the Halting sequence H is defined as [26, Appendix]:

$$I(x; H) = K(x) - K'(x).$$

Proposition 6.1.3.

$$K(x|\underline{y}) - K(x|y) \leqslant_+ I(y; H) + O(\log I(y; H))$$

Proof. Let m be some universal semimeasure with $K(m) \leqslant_+ 0$. Let $t_{l|y}$ and t_k be defined by the conditional variant of Definition 3.1.4. The conditional version of Lemma 3.1.7 shows that for all l $K(t_{y,l}|y) \geqslant^+ l$. For p the witness of $K(x|y)$, let l, k be minimal such that:

$$\begin{aligned} t_{l|y} &\geqslant t[p|y] \\ t_k &\geqslant t[p|y]. \end{aligned}$$

Using the conditional version of Lemma 3.1.7 it follows that

$$t_k \leqslant PBB(k+2\log k+O(1)) \leqslant PBB(k+2\log k+O(1)|y) \leqslant t_{k+2\log k+O(1)|y}.$$

and therefore $l \leqslant k + 2\log k$. Also, remark that since $t_{l-1|y} \leqslant t[p|y] \leqslant t_k$ one has

$$l, k, y, t_k \longrightarrow t_{l-1|y}.$$

Remark that the proposition is formulated independent on the choice of oupi-interpreter ϕ. Therefore a suitable ϕ can be assumed such that

$$K_{t_k}(x|l, k, y, t_k) \leqslant_+ K_{t_k}(x|y, t_{l-1|y}) + O(\log k).$$

This shows for suitable ϕ that

$$\begin{aligned} K(x|y\uparrow) &\leqslant_+ K_{t_k}(x|y, t_k, k, l) + k + O(\log k) \\ &\leqslant_+ K_{t_k}(x|y, t_{l-1|y}) + k + O(\log k) \\ &\leqslant_+ K_{t_k}(x|y) - K(t_{l-1|y}|y) + k + O(\log k) \\ &\leqslant_+ K_{t_k}(x|y) - l + k + O(\log k). \end{aligned}$$

Remark that a program q witnessing $K(t_{l|y}|y) \leqslant_+ l + 2\log l$ in Lemma 6.1.4 shows that

$$k - l \leqslant_+ K(y) - K'(y) + O(\log k).$$

□

Remark that using the result mentioned below Proposition 3.1.20, one can show that k in this proof is lower bounded by $k_c(x, y)$ for some c large enough. This shows that Proposition 6.1.3 is also valid for $\log I(y; H)$ replaced by $\log k_c(x, y)$.

Lemma 6.1.4. *If $\phi_{t_k}(p, y)\downarrow$, then*

$$k - 2\log k - l(p) \leqslant^+ K(y) - K'(y).$$

Proof. Remark that by Lemma 3.1.7 one has $K(t_k, k) \leqslant^+ k + 2\log k$, and that t_k can be computed from k on a Turing machine with Halting oracle, thus $K'(t_k|k) =^+ 0$. Therefore:

$$\begin{aligned} K(y) + l(p) - k &\geqslant^+ K(y, p) - K(t_k) \\ &=^+ K(y, p|t_k^*) \\ &\geqslant^+ K'(y, p|k) \\ &\geqslant^+ K'(y, p) - 2\log k \\ &\geqslant^+ K'(y) - 2\log k. \end{aligned}$$

□

6.1.2 Total sophistication and a total coding result

Remind the definition of total conditional semimeasures P of Definition 5.1.7, defining for each n a constant $P(\varepsilon|\underline{n})$. Also, remind that these semimeasures have a universal Σ-semimeasure denoted by $m(x|\underline{y})$. Remark that an enumeration for $m_t(x|\underline{y})$ of $m(x|\underline{y})$ exists, such that for each fixed t, one has that $m_t(x|\underline{y})$ is total conditional.

Definition 6.1.5. For an universal total length conditional enumeration $m_t(x|\underline{y})$ of a semimeasure y, the *$\underline{m}$-sophistication* is defined by

$$\begin{aligned} t_{k|\underline{n}} &= \min\{t : m(\varepsilon|\underline{n}) - m_t(\varepsilon|\underline{n}) \leqslant 2^{-k}\} \\ k_c(x|\underline{n}) &= \min\{k : K_{\underline{t}_{k|n}}(x|n) \leqslant K(x|n) + c\}. \end{aligned}$$

Remark that $m(\varepsilon|\underline{n})$ now corresponds to the Ω-symbol of Definition 3.1.1. Most propositions for m-sophistication also hold for $\underline{m}$-sophistication. Especially, that the choice of different universal m, maximally results in a logarithmic additive difference. For some total length conditional semimeasure, with a total length conditional enumeration, let $u_c(x|\underline{n})$ be the corresponding $\underline{m}$-sophistication.

Lemma 6.1.6.

$$\forall c \exists c' \Big[abs\big(k_{c-c'}(x|n) - k_c(x|\underline{n})\big) \leqslant O(\log k_c(x|n)) \Big]. \tag{6.1}$$

Proof. Choose

$$\begin{aligned} m_t(x|n) &= \sum\{2^{-l(p)} : \phi_t(p|n) \downarrow = x\} \\ m_t(x|\underline{y}) &= \sum\{2^{-l(p)} : \phi_t(p|\underline{y}) \downarrow = x\}. \end{aligned}$$

Assume the well-defined commands v, w on ϕ such that

- for any $p \in 2^{<\omega}$ with $\phi_t(p|n) \downarrow$: $\phi_t(wp|x, n) \downarrow$ for all $x \in 2^n$,
- for any $p \in 2^{<\omega}$ with $\phi_t(p|y, n) \downarrow \in 2^n$: $\phi_t(vp|n) \downarrow \in 2^n$,

Remark that this shows that for all i

$$m(x|n)^i \longleftrightarrow_+ m(x|\underline{n})^i.$$

This implies for the constructed universal semimeasures that

$$\text{abs}\big(k_c(x|n) - k_c(x|\underline{n})\big) \leqslant O(\log k_c(x|n)).$$

For general semimeasures, the lemma follows by remarking that changing m, changes $k_c(x|n)$ and $k_c(x|\underline{n})$ by an additive logarithmic term. □

Proposition 6.1.7. *For any c large enough*

$$abs\Big(K(x|\underline{y}) + \log m(x|\underline{y})\Big) \leqslant 2\log k_c(x|n)$$

Proof. Let $P_k(x|y)$ be the measure as constructed in equation (3.1) for the universal length conditional semimeasure $m(x|\underline{y})$. In the same way as in the proof of the third item of Proposition 3.1.28, it follows that there is a $k \leqslant k_c(x|\underline{n})$ such that

$$P_k(x|y) \geqslant 2^{-k} m(x|\underline{y}).$$

Also, remark that $K(P_k|n) \leqslant_+ k + 2\log k$. Shannon-Fano code for $P_{k_c(x|\underline{n})}(x|y)$ defines an encoding for x given y. Let $f(p, y)$ be the function that checks whether for any such y, the string p represents a Shannon-Fano code corresponding to the measure $P_{k_c(x|\underline{n})}(x|y)$. Since, P_k is computable, the function f is also computable. Using this function, one shows that

$$K(f) - \log P_k(x|y) \leqslant_+ K(x|\underline{y}).$$

The proposition follows now by remarking that

$$K(f) \leqslant_+ P_{k_c(x|\underline{n})} \leqslant_+ k_c(x|\underline{n}) + 2\log k_c(x|\underline{n}).$$

Applying Lemma 6.1.6, finishes the proof. □

Finally, it is remarked that by Proposition 6.2.7, total conditional complexities define an approximate decomposition of $K(x, y)$.

6.2 Incremental coding

On-line decision complexity has been introduced and investigated in [12, 69]. It also naturally appears in the definition of ideal influence tests as discussed in Chapter 5.

Definition 6.2.1. Let for any n: $x, y \in 2^n$.

- $\phi(p|x\uparrow)\downarrow = y$ is short for:

$$\forall i < n\big[\phi(p|x^i, n)\downarrow = y_{i+1}\big].$$

- The *online complexity* of two strings x, y of equal length is:

$$K(x|y\uparrow) = \min\{l(p) : \phi(p|y\uparrow, n)\downarrow = x\}. \tag{6.2}$$

- The *total online conditional* complexity of a string x given y is:

$$K(x|\underline{y}\uparrow) = \min\big\{l(p) : \forall z \in 2^n[\phi(p, z\uparrow, n)\downarrow] \wedge \phi(p, y\uparrow, n)\downarrow = x\big\}.$$

6.2.1 Coding results

Proposition 6.2.2. *For all computable causal semimeasures P:*

$$-\log P(x|y\uparrow) + K(P) \geqslant^{+} K(x|\underline{y}).$$

Proof. If $E(x|y)$ is the Shannon-Fano code according to $P(x|y)$. We will show that x_{i+1} can be computed from $E(x)$ and y^i. Let $Pc(x|y) = \sum\{P(z|y) : l(z) = n, z < x\}$ according to the construction of the Shannon-Fano code. Let for all $i < n$, $Pc(x^i|y) = Pc(x^i 0...0|y)$. Remark that for $v \in 2^{n-i}$, $Pc(x^{i+1}|y^i v)$ is independent of v, and therefore $Pc(x^{i+1}|y^i)$ is defined and computable. Take $x_{i+1} = 0$ if

$$E(x) \in [Pc(x^i 0|y^i), Pc(x^i 1|y^i)],$$

and $x_{i+1} = 1$ otherwise. □

A coding result for a universal causal semimeasure is now given.

Proposition 6.2.3.

$$-\log m(x|\underline{y}\uparrow) \leqslant_{+} K(x|\underline{y}\uparrow) \leqslant_{+} \log m(x|\underline{y}\uparrow) + O(\log k_0(x|n))$$

Proof. The left inequality follows by considering the semimeasure

$$Q_p(x|\underline{y}\uparrow) = \sum \left\{2^{-l(p)} : \forall z \in 2^n [\phi(p, z\uparrow, n)\downarrow] \wedge \phi(p, y\uparrow, n)\downarrow = x\right\}.$$

Remark now that

$$-\log m(x|\underline{y}\uparrow) \leqslant_{+} -\log Q_p(x|\underline{y}\uparrow) \leqslant_{+} K(x|\underline{y}\uparrow).$$

The right inequality is now shown. Let P_k be defined as in equation (3.1) using $m(x|\underline{y}\uparrow)$, for some $k \leqslant k_0(x|\underline{n})$, as in item (iii) of Proposition 3.1.28. Since P_k is computable, now Proposition 6.2.2 can be applied. Remark that $K(P_k) \leqslant_{+} k + 2\log k$, and that $P_k(x|y) \leqslant m(x|\underline{y}\uparrow)$. □

Remark that this result is complementary to the coding result in [12] for online semimeasures, given below.

Theorem 6.2.4 ([12])**.**

$$-\log m(x|\underline{y}) \leqslant_{+} K(x|y\uparrow) \leqslant_{+} -\log m(x|\underline{y}) + 2\log(-\log m(x|\underline{y})).$$

Remark that online and total online complexity can be very different.

Proposition 6.2.5. *For any n, there exists an $x, y \in 2^n$ such that*

$$\begin{aligned} K(x|y\uparrow) &\leqslant_+ 0 \\ K(x|\underline{y}\uparrow) &\geqslant_+ n/2. \end{aligned}$$

Proof. Let $r \in 2^{n/2}$ such that $K(r|n, a^*) \geqslant n/2$. Let a be the lexicographic first string in $2^{n/2}$ that cannot be computed from r within time $t_{n/2|n}$. Let $y = ar$. Finally, let $x_{n/2+i} = XOR(r_{i-1}, a_i)$ for $n/2 < i < n$, and let all other bits of x be zero. Observe that $K(x|y\uparrow) =_+ 0$, while a total program p computing x from y, also computes a from r, and therefore also computes $t_{n/2|n}$. Therefore, $l(p) \geqslant_+ K(t_{n/2|n}) \geqslant_+ n/2$. □

6.2.2 Decomposition of algorithmic complexity

Additivity of Kolmogorov complexity means that $K(x, y) =^+ K(x) + K(y|x^*)$. This section is devoted to the question whether how one can establish a similar relation using online complexities. Stated somewhat more explicit, by a decomposition of $K(z)$ it is understood that: if a minimal program incrementally computes a part of z, and another minimal program computes the complimentary part of z, the sum of the length of these programs are approximately $K(z)$. Similar to definition 6.2.1 (total) instantaneous causal complexity is defined as:

$$\begin{aligned} K(x|y\uparrow^+) &= \min\{l(p) : i < n, \phi(p, y\uparrow)\downarrow = x\} \\ K(x|\underline{y}\uparrow^+) &= \min\{l(p) : \forall z \in 2^n[\phi(p, z\uparrow)\downarrow] \wedge \phi(p|y\uparrow)\downarrow = x\}. \end{aligned}$$

For strings with constant bounded computational m-sophistication, a nice decomposition

$$K(x|y\uparrow) + K(y|x\uparrow^+) =^+ K(x, y|n)$$

can be easily proved. However, the difference between both terms can be very large for strings with large m-sophistication, as shown in proposition 6.2.6. The same conclusion applies for total incremental conditional complexities, since these are larger than incremental complexities.

Proposition 6.2.6. *There is a constant $c > 0$ such that for all n, there are strings $x, y \in 2^n$ such that:*

$$K(x|y\uparrow) + K(y|x\uparrow^+) - K(x, y|n) \geqslant^+ cn.$$

The proof is technical, and the whole Subsection 6.3 is devoted to it. The main trick is informally explained. First observe that by Theorem 6.3.3, it is possible to construct a sequence of binary sequences z_i, $i \geqslant 0$, for which $K(K(z_i)|z_i)$ is

large. We take as x some concatenation of the z_i (and a little more), and for y some concatenation of the binary expansions of $K(z_i)$ (and a little more), filled with zeros at the right places. It turns out that the information of $K(z_i)$ must be present in both shortest programs. To reuse this information in the second term of the decomposition, we make the decomposition more asymmetric, and limit-computable in stead of co-enumerable. Proposition 6.2.7 shows that within a good approximation the decomposition is valid. Let

$$K(y|\underline{x}\uparrow^+,p) = \min\big\{l(q) : \forall z[\phi(q,z\uparrow^+,p)\downarrow] \wedge \phi(q,x\uparrow^+,p)\downarrow = y\big\}.$$

Proposition 6.2.7. *Let $k_0(x,y|n)$ be the bivariate m-sophistication according to $m(x,y|n)$. Let p be the minimal program in the definition of $K(x|\underline{y})$:*

$$K(x,y|n) \leqslant_+ K(x|\underline{y}) + K(y|p) \leqslant K(x,y|n) + O(\log k_0(x,y|n)).$$

Let p be the minimal program in the definition of $K(\underline{x}|y\uparrow)$:

$$K(x,y|n) \leqslant_+ K(\underline{x}\lceil y\uparrow) + K(y|x\uparrow^+,p) \leqslant K(x,y|n) + O(\log k_0(x,y|n)).$$

Proof. The first equation of the proposition is proved in a similar way as the second one. Therefore the second one is now shown. The $\geqslant^+$ inequality is trivial, therefore it remains to show the $\leqslant^+$ inequality. Let:

$$s \quad = \quad \max\{s : \forall z \in 2^n[\phi(p,z\uparrow)\downarrow\},$$

and let $k = k_0(x,y|n)$. Two cases are now distinguished.
Case $s \leqslant t_k$.
If $s \leqslant t_k$, then

$$K(x,y),k,x,y \longrightarrow x,y,t_k \longrightarrow p$$

and therefore

$$K(p) + K(x,y|p^*) \leqslant^+ K(x,y,p) \leqslant^+ K(x,y,t_k) \leqslant^+ K(x,y) + 2\log k.$$

Because $K(x,y|p) = K(y|\underline{x}\uparrow^+,p)$, this shows that

$$K(x\lceil y\uparrow) + K(y|\underline{x}\uparrow^+,p) =^+ K(x,y).$$

Case $t_k < s$.
Remark that in this case

$$p,x,y \longrightarrow t_k.$$

The result follows now from Lemma 6.2.8. □

Lemma 6.2.8.

$$K(x|y\uparrow) + K(y|\underline{x}\uparrow^+, t_{k_0(x,y|n)}) \leqslant K(x,y) + O(\log k_0(x,y|n)).$$

Proof. Let P_k be the minimal sufficient statistic as in equation (3.1), for a bivariate universal semimeasure, and let $k \leqslant k_0(x,y|n)$ such that $P_k(x,y|n)$ is a c-sufficient statistic as in Proposition 3.1.28. Remark that

$$k, t_{k_0(x,y|n)} \longrightarrow k, t_k \longrightarrow P_k,$$

and

$$-\log P_k(x,y|n) = K(x,y|n) - k + O(\log k). \tag{6.3}$$

Let $P_k(x\lceil y\uparrow)$ and $P_k(y\lceil x\uparrow^+)$ be the causal associated semimeasures of $P_k(x,y|n)$. Remark that

$$-\log P_k(x\lceil|y\uparrow)P_k(y\lceil x\uparrow^+) = -\log P_k(x,y|n), \tag{6.4}$$

and

$$\begin{aligned} K(x|\underline{y}\uparrow) &\leqslant^+ K(P_k) - \log P_k(x\lceil y\uparrow) \\ &\leqslant^+ k + 2\log k - \log P_k(x\lceil y\uparrow) && (6.5)\\ K(y|\underline{x}\uparrow^+, t_{k_0(x,y|n)}) &\leqslant^+ -\log P_k(y\lceil x\uparrow^+). && (6.6) \end{aligned}$$

Adding up Equations (6.3-6.6) finishes the proof. □

Remark that there exist strings for which $k_0(x,y|n)$ is close to $2n$, and in general this provides a logarithmic bound. In the proof two cases are considered, and it is not clear whether they are both essentially possible. If only the first case appears, the $O(\log k_0(x,y|n))$ term can be removed in the theorem.

Question 6.2.9. *Does the equality of Proposition 6.2.7 hold without the $O(\log k_0(xy))$ term ?*

This question can be generalized by remarking that any program p allowed in the definition of $K(x|\underline{y}\uparrow)$, defines for each n, a computable finite set

$$S_p = \{(y, \phi(p|y\uparrow)) : y \in 2^n\}.$$

The set of all S_p for any p that outputs a finite set, defines an enumerable series of finite computable sets. If Question 3.3.15 for $F = \Delta_1$ is answered in the negative, then Question 6.2.9 will be answered in the positive.

Question 6.2.10. *Is following decomposition of $K(x,y)$ valid ?*

$$K(x|y\uparrow) + K(y|x\uparrow^+, p) =^+ K(x,y) + O(\log k_c(x,y))$$

where p is the minimal program in the definition of $K(x|y\uparrow)$.

By a similar argument following Question 6.2.9, if Question 3.3.15 is answered in the negative for $F = \Sigma$, then Question 6.2.10 is answered in the positive, even when the $\log k_c(x,y)$ are removed.

6.2.3 Information transfer and instantaneous common information

As with semimeasures, analogues information transfers can now be defined. Using the decomposition of Proposition 6.2.7, this will lead to a symmetric instantaneous information transfer defined below. A schematically overview of three resulting dependencies between three independent sources and the measured signals are given in figure 6.1. Together, this informally, decomposes $K(x, y)$ in five parts: the information in x if y is given, and in y if x is given, the information that is transferred from the past of x to the future of y and from the past of y to the future of x, and the information that instantly seems to be available from a hypothetical common source, or has been instantaneously transmitted. Such an interpretation is only suitable if the instantaneous mutual information is symmetric. Such a symmetry can follow from a decomposition of $K(x, y|n)$. The first three sum up to the mutual information of x and y. According to the four types of causal complexity, we can introduce four definitions of information transfer.

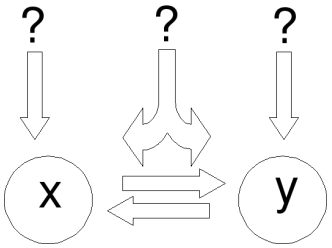


Figure 6.1: Decomposition of information in x and y by three sources. The mutual information as a sum of the three arrows in the middle: information flow from x to y, information flow from y to x and information from a common unknown source, or instantaneous flows.

Definition 6.2.11. Algorithmic, total algorithmic, conditional algorithmic, and total conditional algorithmic information transfer are:

$$\begin{aligned} IT(x \leftarrow y) &= K(x) - K(x|y\uparrow) \\ TIT(x \leftarrow y) &= K(x) - K(x|\underline{y}\uparrow) \\ ITc(y \leftarrow^{+} x) &= K(y) - K(y|x\uparrow^{+}, p) \\ TITc(y \leftarrow^{+} x) &= K(y) - K(y|\underline{x}\uparrow^{+}, p'), \end{aligned}$$

where p and p' correspond to the minimal programs in the definition of $K(x|y\uparrow)$ and $K(x|\underline{y}\uparrow)$.

Instantaneous, total instantaneous, conditional instantaneous and total conditional

instantaneous common information are:

$$\begin{aligned} IT(x\uparrow;y\uparrow) &= K(x|y\uparrow) - K(x|y\uparrow^+) \\ TIT(x\uparrow;y\uparrow) &= K(x|\underline{y}\uparrow) - K(x|\underline{y}\uparrow^+) \\ ITc(x\uparrow;y\uparrow) &= K(x|y\uparrow) - K(x|y\uparrow^+,p) \\ TITc(x\uparrow;y\uparrow) &= K(x|\underline{y}\uparrow) - K(x|\underline{y}\uparrow^+,p), \end{aligned}$$

where p corresponds to the minimal programs in the definition of $K(x|y\uparrow)$ and $K(x|\underline{y}\uparrow)$.

The subsequent corollary follows now from Proposition 6.2.7.

Corollary 6.2.12.

$$TIT(y\leftarrow x) + TIT(x\leftarrow y) + TITc(x\uparrow;y\uparrow) =^+ I(x;y) \pm O(\log k_0(x,y|n)),$$
$$TITc(x;y) = TITc(y;x) \pm O(\log k_0(x,y|n)).$$

Question 6.2.13. *How symmetric is $ITc(x\uparrow;y\uparrow)$?*

In the case the universal prefix-free Turing machine has tapes with symbols from a large finite alphabet, it can be easily shown from Proposition 6.2.6, that $IT(x\uparrow;y\uparrow)$ and $TIT(x\uparrow;y\uparrow)$ are not symmetric. Take an alphabet of 2^m symbols and let $x_i = z^{(i)}$ and let $y_i = K(z^{(i)})$. Then:

$$I(x\uparrow;y\uparrow) - I(y\uparrow;x\uparrow) > o(\log n).$$

6.3 Additivity of online complexity is violated

It is shown that for any n, there a $z \in 2^n$, such that the sum of the online complexity of predicting the even bits of z given the previous uneven bits, and the online complexity of predicting the uneven bits given the previous even bits, exceeds the Kolmogorov complexity of z by a linear term in the length of z. It is also shown that instantaneous mutual information of two sequences in $\omega^{<\omega}$, is asymmetric by a linear constant in the length of some natural encoding of the sequences.

6.3.1 Muchnik's paradox

First some additional motivation based on martingales, and an alternate proof strategy are discussed.

Suppose the following bets on a random variable X in $2^{<\omega}$. An agents pays one unit of utility, subsequently, he chooses some n and $x \in 2^n$, then X^n is observed, and if $x = X^n$, a reward of $2^n M(x)$ utility is paid by the banker. When M satisfies $M(\varepsilon) \leqslant 1$, and for any $x \in 2^{<\omega}$ satisfies:

$$m(x) \geqslant \frac{1}{2}(m(x0) + m(x1)), \tag{6.7}$$

a booker who beliefs that the variable X is distributed uniformly on 2^∞, will agree to be the banker in such a bet. A *supermartingale* is a positive function M on $2^{<\omega}$ with $M(\varepsilon) \leqslant 1$ such that equation (6.7) is satisfied. A sequence $\alpha \in 2^\infty$ *succeeds* on a martingale M, iff $M(\alpha^i)$ is unbounded for increasing i. A sequence α is Martin-Löf random if for any Σ-martingale, α is not a winning sequence [1]. This means that when an agent knows that α is Martin-Löf random, he may maximally loose a finite amount of utility, when he plays the banker in a game where the rewards correspond to a Σ-martingale. Remark that the reward function is a Σ-function. This means that the bank pays a reward distributed through time.

For $\beta \in 2^\infty$, it is said that α is *Martin-Löf random given* β iff no martingale that is lower semicomputable given β succeeds on α. α is said to be *online Martin-Löf* given β iff no martingale M succeeds on α, such that for any i, the value $M(\alpha^i)$ can be evaluated from β^i. The van Lambalgen theorem implies that α is Martin-Löf random iff $\alpha_{0,2,...}$ is Martin-Löf random given $\alpha_{1,3,...}$ and vice versa. The question rises whether it also holds that α is Martin-Löf random iff $\alpha_{0,2,...}$ is online Martin-Löf random given $\alpha_{1,3,...}$ and vice versa.

Muchnik's paradox [47] states that there are sequences α that succeed on a lower semicomputable martingale, but do not succeed on any even and uneven online lower semicomputable martingales. In other words, if a booker knows that it is safe to be the banker for some game on even bits of a sequence, and he also knows that it is safe for some game on the uneven bits, then it is still not safe to be the banker for a game on all bits of a sequence. Remark that this can only happen for reward functions in $\Sigma \backslash \Delta_1$.

The proof presented in [47], can also be used to show that the online complexities of predicting even and uneven bits differs by a linear constant from $K(x)$. This is possible by the online Coding Theorem, and the simple adaptation of keeping the h parameter constant. Even more, this proof is substantially more optimal when expressed in number of pages, and dependencies on other results, then the proof given below. However, here it will also be shown that online complexity is asymmetric by a linear term in the length of a binary sequence, which does not follow from Muchnik's proof. On one side the two approaches allow to show the non-additivity of online Kolmogorov complexity, on the other side, they also have each very unrelated consequences, which seem to be far from equivalent. Therefore, both proofs of non-additivity of online Kolmogorov complexity seem to exploit two different structures.

[1] In many textbooks Martin-Löf randomness is characterized differently, and an equivalence with this definition is shown.

6.3.2 Main result and proof tactic

For any $x \in 2^{<\omega}$, $l(x)$ denotes the length of x. For any $x \in \omega^{<\omega}$, $l(\overline{x})$ corresponds to the length of some prefix-free encoding of x on a binary tape:

$$l(\overline{x}) = \sum_{i=1}^{l(x)} 2 \log x_i.$$

Proposition 6.3.1.

$$\exists c > 0 \exists^{\infty} x, y \in \omega^{<\omega} \\ \big[K(x|y\uparrow) + K(y|x\uparrow^{+}) - K(x,y) > c(l(\overline{x}) + l(\overline{y}))\big].$$

Proposition 6.3.2.

$$\exists c > 0 \exists^{\infty} x, y \in \omega^{<\omega} \\ \big[IT(x\uparrow; y\uparrow) - IT(y\uparrow; x\uparrow) > c(l(\overline{x}) + l(\overline{y}))\big].$$

In [23] and repeated in [25, 44], it is proven the complexity of complexity can be high. The result also follows for the conditioned case.

Theorem 6.3.3. *For any $n \in \omega$, and $w \in \omega^{<\omega}$ there is an $a \in 2^n$*

$$K(K(a|w)|a, w) \geqslant_{+} \log n - \log\log n.$$

Let y be the binary expansion of $K(x)$. It follows that

$$K(x) + K(y|x) - K(x,y) \geqslant_{+} \log n - \log\log n.$$

By inserting zeros at the right places in x, y, it can be shown that there exists infinity many $x, y \in 2^n$:

$$K(x|y\uparrow) + K(y|x\uparrow^{+}) - K(x,y) > O(\log n).$$

This shows proposition 6.3.1 for a logarithmic term in $l(x)$. It seems natural to think that this logarithmic difference can be improved to a linear difference, by concatenating a sufficiently large amount of such strings. However, to be able to add up these differences, conditional complexities must add up in some way to on-line Kolmogorov complexity, which is not possible within sufficient accuracy.

The main trick to solve this issue, lies in Lemma 6.3.4, which shows additivity of conditional complexities to online complexity, if some additional information is available. This information is given by the numbers $L_{X,i}$, representing the amount of programs of length N, that are algorithmic solutions for a series of subtasks. By

using the Coding Theorem, and the numbers $L_{X,i}$, the conditional Kolmogorov complexities, relates to the online complexities.

The additional information is available in the sequences u and v, and is combined with x and y. From these observations Proposition 6.3.1 is shown. To show Proposition 6.3.2, remark that for any $a \in 2^{<\omega}$ one has

$$K(K(a)) + K(a|K(a)^*) =_+ K(K(a), a) =_+ K(a),$$

by additivity of K and by Theorem 2.4.7. By exploiting this property, it is shown that additivity of online complexity in the other direction is satisfied.

6.3.3 Proof

The proof below uses Kolmogorov complexities of multiple sequences in $\omega^{<\omega}$. Strict mathematically, such complexities are not defined, however, by assuming a suitable computable bijections between $(\omega^{<\omega})^{\omega}$ and $\omega^{<\omega}$, one can assume that such complexities as defined.

A *task* $T = (A \to Q)$ is defined by two series A, Q of equal length, of $\omega^{<\omega}$-tuples. For some task T of length n, the set of solutions of T given N, with length N is defined by

$$S_{T,N} = \big\{p \in 2^N : \forall i < n \left[\phi(p|Q^i, N) \downarrow = A_i\right] \big\}.$$

The log-cardinality is denoted as

$$L_{T,N} = \log |S_{T,N}|.$$

Let

$$K(T) = \min \big\{p \in 2^N : \forall i < n \left[\phi(p|Q^i) \downarrow = A_i\right] \big\}.$$

Let $T_i = (Q^i, A^{i-1} \to A_i)$ and $T^i = (Q^i \to A^i)$. A bound for $K(T)$ is proven for a task T of length n.

Lemma 6.3.4.

$$K(T) \geqslant \min \left\{ N, \sum_i K(T_i | L_{T^{i-1},N}) - O(n) \right\}.$$

Proof. If $|S_{T,N}| = 0$, then no string of length N can solve task T, thus $K(T) > N$, and the Lemma is proven. Suppose that $|S_n| \geqslant 1$. Let $L_i = L_{T^i,N}$. For each i, a semimeasure P can be constructed using A^{i-1},Q^i,L_{i-1},N:

$$P(z) = 2^{-L_{i-1}} \Big| \left\{p \in S_{T^{i-1},N} : \phi(p|Q^i, N) \downarrow = z\right\} \Big|.$$

Remark that P defines a Σ-semimeasure given A^{i-1}, Q^i, L_{i-1}, N. Since $|S_{T^i,N}| \geqslant |S_{T,N}| \geqslant 1$, it follows that $P(A_i) > 0$, for $i \leqslant n$. Applying the Coding Theorem [44] on P, it follows that:

$$L_{i-1} - L_i \geqslant K(T_i|L_{i-1}, N) - O(1).$$

Summing over i:

$$L_0 - L_n \geqslant \sum_i K(T_i|L_{i-1}, N) - O(n). \tag{6.8}$$

Let p be a program of length $K(T)$, solving task T. It possible to append $2^{N-K(T)-O(1)}$ different strings of length $N - K(T) - O(1)$ to p, in order to obtain elements from S_n. Therefore:

$$L_n \leqslant^+ N - K(T). \tag{6.9}$$

Observe that $L_0 = N$. Combining equations (6.8) and (6.9) proves the lemma. □

Proof of Propositions 6.3.1.
For some n, let $u, x, y, v \in \omega^n$, and let $N = 2mn$ for some m large enough.

$$z = u_0 x_0 y_0 v_0 \ldots u_n x_n y_n v_n.$$

The following tasks will be used in the proof:

$$\begin{aligned} T_{ux,i} &= (z^{4(i-1)}, N \to u_i, x_i) \\ T_{x,i} &= (z^{4(i-1)+1}, N \to x_i) \\ T_{yv,i} &= (z^{4(i-1)+2}, N \to y_i, v_i). \end{aligned}$$

u, x, y, v will be constructed to satisfy

$$K(u, x, y, v|N) \leqslant K(T_{ux}) + K(T_{yv}) - n \log m + O(n \log \log m). \tag{6.10}$$

To show that the proposition follows from this equation, let $\langle a, b\rangle$ be the bijective computable pairing function $\langle a, b\rangle = \frac{1}{2}(a + b)(a + b + 1) + a$. Remark that $l(a) + l(b) \geqslant_+ l(\langle a, b\rangle)$. The proposition follows for

$$x'_i = \langle u_i, x_i\rangle \tag{6.11}$$

$$y'_i = \langle y_i, v_i\rangle, \tag{6.12}$$

since equation (6.14) shows that $l(\overline{x'}) + l(\overline{y'}) \leqslant O(mn)$, and by substituting

$$\begin{aligned} K(x'|y' \uparrow) &\leqslant_+ K(T_{ux}) + K(N) \\ &\leqslant_+ K(T_{ux}) + 2 \log mn \\ K(y'|x' \uparrow^+) &\leqslant_+ K(T_{yv}) + K(N) \\ &\leqslant_+ K(T_{yv}) + 2 \log mn \end{aligned}$$

in equation (6.10).

It remains to construct u, x, y, v for $N = 2mn$ such that equation 6.10 is valid. For $X \in \{ux, yv\}$ and $i < n$ the short notation $L_{X,i} = L_{T_X^i,N}$ is used. Also, let $D_{X,0} = 0$, and for $1 \leqslant i < n$ let

$$D_{X,i} = L_{X,i-1} - L_{X,i}.$$

Remark that

$$\sum_{j \leqslant i} D_{X,i} = N - L_{X,i},$$

and thus

$$D_X^i \longrightarrow L_{X,i}.$$

The inequalities (6.13)-(6.18) are now derived.

1. Let

$$\begin{aligned} u_i &= D_{yv,i-1} \\ v_i &= D_{ux,i}. \end{aligned}$$

Remark that

$$\begin{aligned} \sum_{i<n} u_i &\leqslant L_{yv,0} = N \\ \sum_{i<n} v_i &\leqslant L_{ux,0} = N. \end{aligned}$$

Since $N = 2mn$ and by concavity of the $\log$ function it follows that

$$\begin{aligned} &\sum_{i<n} \log^{(3)} u_i \leqslant n \log^{(3)} m \\ &\sum_{i<n} \log^{(3)} v_i \leqslant n \log^{(3)} m. \end{aligned} \tag{6.13}$$

This also implies for m, n large enough that

$$\sum_{i<n} 2(l(\overline{u}) + l(\overline{x}) + l(\overline{y}) + l(\overline{v})) \leqslant 2mn = N. \tag{6.14}$$

2. Remark that

$$\begin{array}{ccccccc} z^{4(i-1)}, N & \longrightarrow & v^{i-1}, N & \longrightarrow & D_{ux}^{i-1}, N & \longrightarrow & L_{ux,i-1} \\ z^{4(i-1)+2}, N & \longrightarrow & u^i, N & \longrightarrow & D_{yv}^{i-1}, N & \longrightarrow & L_{yv,i-1}. \end{array}$$

Therefore, for $X \in \{ux, yv\}$, and by equation (6.14), Lemma 6.3.4 implies

$$K(T_X) \geqslant \sum_{i<n} K(T_{X,i}). \tag{6.15}$$

3. Let
$$y_i = K(T_{x,i}) = K(x_i|z^{4(i-1)+1}, N).$$
By Corollary 2.4.8 it follows that
$$\begin{aligned}
&K(u_i, x_i|z^{4(i-1)}, N)\\
&\geqslant_+ K(u_i, x_i, K(x_i|u^i, K(u^i|z^{4(i-1)}, N), z^{4(i-1)}, N)|z^{4(i-1)}, N)\\
&\geqslant_+ K(u_i, x_i, K(x_i|u^i, z^{4(i-1)}|z^{4(i-1)}, N) - 2\log^{(3)} u_i.
\end{aligned}$$
Since $K(x_i|u_i, z^{4(i-1)}) =_+ K(x_i|z^{4(i-1)+1}) = y_i$, one has
$$K(u_i, x_i|z^{4(i-1)}) \geqslant_+ K(u_i, x_i, y_i|z^{4(i-1)}) - 2\log^{(3)} u_i. \tag{6.16}$$

4. Let $m \in \omega$, large enough. By Theorem 6.3.3 for each i, one can choose $x_i \leqslant 2^m$ such that
$$K(K(x_i|z^{4(i-1)+1}, N)|x_i, z^{4(i-1)+1}, N) \geqslant_+ \log m - \log\log m.$$
This shows that
$$\begin{aligned}
K(y_i|z^{4(i-1)+2}, N) &=_+ K(K(T_{x,i})|z^{4(i-1)+2}, N)\\
&=_+ K(K(x_i|z^{4(i-1)+1}, N)|x_i, z^{4(i-1)+1}, N)\\
&\geqslant_+ \log m - \log\log m.
\end{aligned} \tag{6.17}$$

5. By additivity of K
$$\begin{aligned}
K(y_i, v_i|z^{4(i-1)+2}, N) &=^+ K(y_i|z^{4(i-1)+2}, N)\\
&\quad +K(v_i|y_i^*, z^{4(i-1)+2}, N)\\
&\geqslant^+ K(y_i|z^{4(i-1)+2}, N)\\
&\quad +K(v_i|y_i, z^{4(i-1)+2}) - \log\log y_i.
\end{aligned}$$
Since $y = K(T_{x,i}) \leqslant 2l(x) = 2m$, one has
$$\begin{aligned}
K(v_i|z^{4(i-1)+3}, N) &\leqslant_+ K(y_i, v_i|z^{4(i-1)+2}, N)\\
&\quad -K(y_i|z^{4(i-1)+2}, N)\\
&\quad +2\log\log m.
\end{aligned} \tag{6.18}$$

Combining equations (6.15)-(6.13) shows equation (6.10):
$$\begin{aligned}
&K(u, x, y, v|N)\\
&\leqslant \textstyle\sum_{i<n} K(u_i, x_i, y_i|z^{4(i-1)}, N)\\
&+ \textstyle\sum_{i<n} K(v_i|z^{4(i-1)+3}, N) + O(n)\\
&\leqslant \textstyle\sum_{i<n} K(u_i, x_i|z^{4(i-1)}, N)\\
&+ \textstyle\sum_{i<n} 2\log^{(3)} u_i && \text{by (6.16)}\\
&+ \textstyle\sum_{i<n} K(y_i, v_i|z^{4(i-1)+2}, N)\\
&- \textstyle\sum_{i<n} K(y_i|z^{4(i-1)+2}, N) + O(n\log\log m) && \text{by (6.18)}\\
&\leqslant \textstyle\sum_{i<n} K(T_{uv,i}) + \sum_{i<n} K(T_{yv}) && \text{by (6.15)}\\
&-n\log m + O(n\log\log m) && \text{by (6.17),(6.13).}
\end{aligned}$$

□

The proof is now continued.

Proof of Proposition 6.3.2. The proof is shown when $x, y \in \omega^{<\omega}$. The case where $x, y \in 2^{<\omega}$ by using the binary expansions of x, y, and adding zeros where necessary.

Let u, x, y, v and n, m, N as constructed in the proof of Proposition 6.3.1. Assume x', y' as in equations (6.11),(6.12). Let

$$\begin{aligned} T_{y'} &= (z^{4(i-1)}, N \to y'_i) \\ T_{x'} &= (z^{4(i-1)}, y'_i, N \to x'_i). \end{aligned}$$

By equation (6.10), it remains to show that

$$K(T_{y'}) + K(T_{x'}) \leqslant K(x', y') + O(n \log\log m). \tag{6.19}$$

Remark that

$$\begin{aligned} &\sum_{i<n} K(K(y'_i|z^{4(i-1)}, N)|y_i, z^{4(i-1)}, N) \\ &\leqslant \sum_{i<n} 2\log\log y'_i \\ &\leqslant \sum_{i<n} 2\log\log y_i + 2\log\log v_i \\ &\leqslant O(n\log\log m). \end{aligned}$$

Equation (6.19) follows now:

$$\begin{aligned} &K(T_{y'}) + K(T_{x'}) \\ &\leqslant \textstyle\sum_{i<n} K(y'_i|z^{4(i-1)}, N) + \sum_{i<n} K(x'_i|y'^i, z^{4(i-1)}, N) + O(n) \\ &\leqslant \textstyle\sum_{i<n} K(y'_i|z^{4(i-1)}, N) \\ &\quad + \textstyle\sum_{i<n} K(x'_i|y'^{i-1}, K(y'_i|z^{4(i-1)}, N), z^{4(i-1)}, N) \\ &\quad + \textstyle\sum_{i<n} 2\log K(y'_i|z^{4(i-1)}, N) + O(n) \\ &\leqslant \textstyle\sum_{i<n} K(x'_i, y'_i|z^{4(i-1)}, N) + O(n\log\log m) \\ &\leqslant K(x', y'|N) + O(n\log\log m), \end{aligned}$$

by additivity of K. □

Corollary 6.3.5. *There are infinitely many $x, y \in \omega^{<\omega}$ such that*

$$K(x|\underline{y}\uparrow) + K(y|\underline{x}\uparrow) - K(x, y) \geqslant o(l(\overline{x}) + l(\overline{y})).$$

Proof. Direct, since $K(x|\underline{y}) \geqslant K(x|y)$. □

Corollary 6.3.6. *For some $c > 0$, for all but finitely many n, there exist a $z \in 2^{2n}$ such that:*

$$K(0 \to z_1; ...; z_{2n-2} \to z_{2n-1}) + K(z_1 \to z_2; ...; z_{2n-1} \to z_{2n}) - K(z) \geqslant cn.$$

Proof. Let x', y' be as constructed in the proof of Proposition 6.2.6. Let $\overline{x'_i}$ and $\overline{y'_i}$ be binary prefix-free encodings corresponding to the definition of $l(\overline{x})$. Define z such that online complexities correspond to the complexities of the corollary by adding zeros at some places:

$$\begin{aligned} z = \quad & \overline{x'}_{1,1}, 0, ..., \overline{x'}_{1,l(\overline{x'_1})}, 0, \\ & 0, \overline{y'}_{1,1}, ..., 0, \overline{y'}_{1,l(\overline{y'_1})}, \\ & \quad ... \\ & \overline{x'}_{n,1}, 0, ..., \overline{x'}_{n,l(\overline{x'_n})}, 0, \\ & 0, \overline{y'}_{n,1}, ..., 0, \overline{y'}_{n,l(\overline{y'_n})}. \end{aligned}$$

Since $\sum_{i \leqslant n} l(\overline{x'_i}) + l(\overline{y'_i}) \leqslant 3mn$, we have that $z \in 2^{\leqslant 6n}$. This shows that for all but finitely many n a string of length maximally $6mn$ exists that satisfies the inequality of the lemma. By appending zeros to the end of x' and y', the conditions of the corollary can be satisfied for every n. □

7 Conclusions

Many hypotheses, such as the general hypotheses of independence and influence, are not simple, and statistical tests can not be directly defined using tools such as ratio testing or by constructing a uniform most powerful tests. However, when statistics and computability theory are connected one can define conceptual solutions for these problems, using universal lower semicomputable semimeasures, and subsequently apply techniques of hypothesis testing to these tests. Similar conclusions as for simple hypotheses can now be obtained as for such tests. Two techniques from simple hypotheses testing are used for this purpose: sumtests, which are an abstract model for significance testing, and ratio tests, which are connected to Bayesian belief theory.

The search for non-trivial and large sumtests shows that next to the class of lower semicomputable sumtests, also the class of upper semicomputable sumtests should be investigated. This is especially true for tests of independence, where the first class contains only trivial elements. Many simple questions can be formulated, such as which of the two classes has the largest elements, and how large can these elements be. While these questions are simple and have a natural motivation, there answers show a deep structure within the theory of algorithmic information. Only a few of these questions have been addressed in this manuscript.

To define lower semicomputable universal semimeasures corresponding to the hypotheses of time series being influence free, more care must be performed, and therefore the formalism of the definition of objective probabilities, within the computability framework is given in detail. This results in several ratio tests for hypotheses on causality and influence in time series. Comparison of these tests,

shows the remarkable property that there exists time series for which a plausible direction of instantaneous information flow can by determined without direct reference to external information. Another approach to define tests for causality and influence, is given by Bayesianally associate a causal semimeasures with conditional semimeasures. It is shown that the first class of causal semimeasure is disjunct with the causal semimeasures associated with a universal semimeasure, however, how large these differences can be, seems to be a difficult and deep question.

Since it is very difficult to approximate all these tests directly from there definitions, characterizations of the tests are investigated using Kolmogorov complexities. Such characterizations allow us to use data compression heuristics to define more practical tests. All tests investigated in this manuscript have been shown to have such characterizations. For the influence tests this leads to many interesting notions of conditional Kolmogorov complexity, such as total conditional complexities and several online complexities. Traditional Kolmogorov complexity satisfies an additivity property, which implies the necessary symmetry to interpret mutual information as an independence test. For total online complexity, such a decomposition can be defined, and the corresponding instantaneous mutual information is symmetric. However, for the non-total case this is an open question. Furthermore, the tests obtained conceptually have interesting connections with actual tests for influence.

Finally, minimal typical models and minimal sufficient statistics have been investigated. An explicit minimal typical model has been constructed, and it has been shown that it is equivalent with the introduced minimal weak sufficient statistic, but can be very different from the minimal sufficient statistic for very large and complex strings. The description length of this minimal typical model has been called m-sophistication. This m-sophistication appears almost in any non-trivial problem in this theses.

In conclusion, we remark that there is a large amount of publications and textbooks on the application of computability concepts on the definitions of randomness for infinite strings, however, there are only a few publications studying statistical tests for finite strings. This work shows that simple questions originating from the study of the interpretation of probability and statistical hypotheses testing, directly lead to mathematical definitions with a rich structure, and also raise many non-trivial open questions. Additionally, due to many coding theorems and similarities with existing algorithms used in practice, these mathematical observations, have the potential to broaden and formalize some intuitions used by engineers to develop better tests.

Bibliography

[1] L. Antunes. *Useful information*. PhD thesis, Universidade do Porto, 2002.

[2] L. Antunes, A. Costa, A. Souto, and P. Vitanyi. Computation and logic in the real world. In *Computability in Europe*, 2007.

[3] L. Antunes and L. Fortnow. Sophistication revisited. *Theor. Comp. Sys.*, 45(1):150–161, 2009.

[4] L. Antunes, A. Matos, A. Souto, and P. Vitanyi. Depth as randomness deficiency. *Theor. of Comput. Sys.*, 45(4):724–739, 2009.

[5] B. Bauwens. Co-enumerable sumtests for the universal distribution. Submitted, 2008.

[6] B. Bauwens. On the equivalence between minimal sufficient statistics, minimal typical models and initial segments of the Halting sequence. *ArXiv e-prints*, November 2009. `http://arxiv.org/abs/0911.4521`.

[7] B. Bauwens and S. Terwijn. Notes on sum-tests and independence tests. Accepted for publication in Theor. Comput. Sys., open access, 2009.

[8] B. Bauwens, B. Wyns, D. Devlaminck, G. Otte, L. Boullart, and P. Santens. Measuring instantaneous directed dependencies in interacting oscillators. In *Proceedings of the 28th Symposium on Information Theory in the Benelux*, 2007. http://www.autoctrl.ugent.be/bruno/papers/benelux.pdf.

[9] C.S. Calude, P.H.Hertling, B. Khoussainov, and Y. Wang. Recursively enumerable reals and chaitin ω numbers. *Theoretical Computer Science*, 255:125149, 2001.

[10] C. M. Caves. Probabilities as betting odds and the dutch book, June 2000.

[11] G.J. Chaitin. A theory of program size formally identical to information theory. *J. Assoc. Comput. Mach.*, 22(3):329–340, 1975.

[12] A. Chernov, S. Alexander, N. Vereshchagin, and V.Vovk. On-line probability, complexity and randomness. In *ALT '08: Proceedings of the 19th international conference on Algorithmic Learning Theory*, pages 138–153, Berlin, Heidelberg, 2008. Springer-Verlag.

[13] J. D. Collier. *Causation and Laws of Nature*, chapter Causation is the transfer of information, pages 215–246. Kluwer A. Publishers, 1999.

[14] T. M. Cover. *The Impact of Processing Techniques on Communications.*, chapter Kolmogorov Complexity, Data Compression and Inference., pages 23–33. J. Skwyrzynski, Martinus Nijhoff Publishers, 1985.

[15] A. R. Day. Increasing the gap between descriptional complexity and algorithmic probability. *Annual IEEE Conference on Computational Complexity*, 0:263–273, 2009.

[16] S. de Rooij and P.M.B. Vitányi. Approximating rate-distortion graphs of individual data: experiments in lossy compression and denoising. Submitted.

[17] M. Ding, Y. Chen, and S. L. Bressler. Granger causality: basic theory and application to neuroscience. In M. Winterhalder and J. Timmer, editors, *Handbook of Time Series Analysis*, pages 437–460. Wiley-cvh, Wienheim, 2006.

[18] R. Downey and D. Hirschfeldt. Algorithmic randomness and complexity. To appear.

[19] M. Feder and N. Merhav. Universal composite hypothesis testing: a competitive minimax approach. *IEEE Transactions on information theory*, 48(6):1504–1517, 2002.

[20] U. Feldmann and J. Bhattacharya. Predictability improvement as an asymmetrical measure of interdependence in bivariate time series. *International Journal of Bifurcation and Chaos*, 14(2):505–514, 2004.

[21] B. De Finetti. La prvision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincar*, 7:1–68, 1937. Translated as "Foresight. Its logical laws, its subjective sources", in *Studies in Subjective Probability*, H.e. Kyburg, Jr. and H.E. Smokler, Robert E. Krieger Publiching Company, 1980.

[22] B. De Finetti, A. Machi, and A. Smith. *Theory of Probability: A Critical Introductory Treatment*. Wiley, New York, 1993.

[23] P. Gács. On the symmetry of algorithmic information. *Soviet Math. Dokl.*, 15:1477–1480, 1974.

[24] P. Gács. On the relation between descriptional complexity and algorithmic probability. *Theor. Comput. Sci.*, 22:71–93, 1983.

[25] P. Gács. Lecture notes on descriptional complexity and randomness. Technical report, Comput. Sci. Dept., Boston, 1988-2010. Technical report, `http://www.cs.bu.edu/faculty/gacs/papers/ait-notes.pdf`.

[26] P. Gács, J. Tromp, and P.M.B. Vitányi. Algorithmic statistics. *IEEE Trans. Inform. Theory*, 47(6):2443–2463, 2001.

[27] C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

[28] P.D. Grunwald, I.J.Myung, and M.A.Pitt. *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2005.

[29] A. Hajek. Dutch book arguments. In *The Handbook of Rational and Social Choice*, pages 173–196. Oxford University Press, 2009.

[30] M. Hutter. Optimality of universal Bayesian prediction for general loss and alphabet. *Journal of Machine Learning Research*, 4:971–997, 2003.

[31] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004. 300 pages, http://www.idsia.ch/∼marcus/ai/uaibook.htm.

[32] M. Hutter and An. A. Muchnik. Universal convergence of semimeasures on individual random sequences. In S. Ben-David, J. Case, and A. Maruoka, editors, *Proc. 15th International Conf. on Algorithmic Learning Theory (ALT-2004)*, volume 3244 of *LNAI*, pages 234–248, Padova, 2004. Springer, Berlin.

[33] J.M.Keynes. *A Treatise on Probability*. Macmillan and Co, 1921.

[34] G. John. Inference and causality in economic time series models. In Z. Griliches and M. D. Intriligator, editors, *Handbook of Econometrics*, volume 2, chapter 19, pages 1101–1144. Elsevier, 1984.

[35] D.W. Juedes, J.I. Lathrop, and J.H. Lutz. Computational depth and reducibility. *Theor. Comput. Sci.*, 132(1-2):37–70, 1994.

[36] M. Kaminski, M. Ding, W.A. Truccolo, and S.L. Bressler. Evaluating causal relations in neural systems: granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics*, 85(2):145–157, August 2001.

[37] S. C. Kleene. General recursive functions of natural numbers. *Math. Ann.*, 112:727742, 1936.

[38] A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. English translation by Hathan Morrison in *Foundations of the Theory of Probability*, 1950, New York.

[39] M. Koppel. *The Universal Turing Machine: A Half-Century Survey*, chapter Structure, pages 435–452. R. Herken, Oxford University Press, 1988.

[40] P.S. Laplace. *A Philosophical Essay on Probabilities*. Dover Publications Inc., New York, 1951. Translation of a manuscript from 1814.

[41] R.J. Lawrence, S.A. Mulaik, and J.M. Brett. *Causal Analysis*. Sage Publications, 1983.

[42] J. Lemeire, K. Steenhaut, and A. Touhafi. When are graphical causal models not good models? In J. Williamson, F. Russo, and P. McKay, editors, *Causality in the Sciences*. Oxford University Press, 2010.

[43] L.A. Levin. Randomness conservation inequalities; information and independence in mathematical theories. *Inf. Control*, 61(1):15–37, 1984.

[44] M. Li and P.M.B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York, 2008.

[45] I. Martel. *Probabilistic Empiricism: In Defence of a Reichenbachian Theory of Causation and the Direction of Time*. PhD thesis, University of Colorado, 2000.

[46] M.Blum. A machine-independent theory of the complexity of recursive functions. *Journal of the ACM*, 14(2):322–336, April 1967.

[47] A. Muchnik. Algorithmic randomness and splitting of supermartingales, 2008.

[48] A. Nies. *Computability and Randomness*. Oxford University Press, Inc., New York, 2009.

[49] N.K.Verschagin. Agorithmic minimal sufficient statistics: a new definition. Presented on the 4th conference on randomness, computability and logic, Luminy., nov 2009.

[50] P. Odifreddy. *Classical Recursion Theory, Volume I*. North-Holland, 1989.

[51] P. Odifreddy. *Classical Recursion Theory, Volume II*. North-Holland, 1999.

[52] M. Palus and A. Stefanovska. Direction of coupling from phases of interacting oscillators: an information theoretic approach. *Physical Review E, Rapid Communications*, 67:055201(R), 2003.

[53] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[54] M. G. Rosenblum and A. S. Pikovsky. Detecting direction of coupling in interacting oscillators. *Phys. Rev. E*, 64(4):045202, Sep 2001.

[55] B. Ryabko, J. Astola, and A. Gammerman. Application of kolmogorov complexity and universal codes to identity testing and nonparametric testing of serial independence for time series. *Theoretical Computer Science*, 359(1-3):440–448, august 2006.

[56] K. Sameshima and L.A. Baccala. Using partial directed coherence to describe neuronal ensemble interactions. *Journal of Neuroscience Methods*, 94:93–103(11), Dec 1999.

[57] T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, Jul 2000.

[58] C.E. Shannon. A mathematical theory of communication, Jul and Oct 1948.

[59] A. Shen. Unpublished work of a.a. muchnik, 2009. Talk on conference Logic Computability and Randomness.

[60] Ph. Smets. No Dutch book can be built against the TBM even though update is not obtained by bayes rule of conditioning. In *SIS, Workshop on Probabilistic Expert Systems*, pages 181–204, 1993.

[61] D. A. Smirnov and B. P. Bezruchko. Estimation of interaction strength and direction from short and noisy time series. *Phys. Rev. E*, 68(4):046209, Oct 2003.

[62] R. I. Soare. *Recursively enumerable sets and degrees*. Springer-Verlag, 1987.

[63] R.J. Solmonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Trans. on Inform. Theor.*, 24(4), july 1978.

[64] R.J. Solomonoff. The probability of "undefined" (non-converging) output in generating the universal probability distribution. *Information Processing Letters*, 106(6):238 – 240, 2008.

[65] R.M. Solovay. Draft of a paper (or series of papers) on chaitin's work ... done for the most part during the period of sept. - dec. 1974. 215 pp., May 1975.

[66] P. Suppes. *A Probabilistic Theory of Causality*. North-Holland Publishing Company, 1970.

[67] S. A. Terwijn. *Syllabus Computability Theory*. PhD thesis, Technical University of Vienna, 2008.

[68] A.M. Turing. On computable numbers with an application to the entscheidungsproblem. *Proc. London Math. Soc.*, 42:230–265, 1937. corrections ibid. 43 (1937) 544-546.

[69] V.A. Uspensky and S. Alexander. Relations between varieties of kolmogorov complexities. *Theory of Computing Systems*, 29:271–292, 1996.

[70] N.K. Vereshchagin and P.M.B. Vitányi. Kolmogorov's structure functions and model selection. *IEEE Trans. Infor. Theory*, 50(12):3265–3290, 2004.

[71] P.M.B. Vitányi. Meaningful information. *IEEE Trans. Inform. Theory*, 52(10):4617–4626, 2006.

[72] R. von Mises. *Wahrscheinlichkeitsrechnung und Statistik und Wahreit*. Macmillan ?, New York ?, 1928. revised English edition: *Probability, Statistics, and Truth*, 1939 and 1957.

[73] V. Vovk and G. Shafer. Kolmogorov's contributions to the foundations of probability. *Problems of INformation Transmission*, 39(1):21–31, 2003. `www.probabilityandfinance.com`.

[74] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistritz, R. Bauer, J. Timmer, and H. Witte. Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. *Signal Processing*, 85:2137–2160, 2005.